REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)* 31-12-2016	**2. REPORT TYPE** Final	**3. DATES COVERED** *(From - To)* Jul 1 2015 - Dec 31 2016

4. TITLE AND SUBTITLE

Leveraging Small-Lexicon Language Models

5a. CONTRACT NUMBER
HR0011-15-C-0117

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Cooper, Doug

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

CRCL INC
820 Calle Pluma
San Clemente, California
92673

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Defense Advanced Research Projects Agency
Information Innovation Office (I2O)
675 North Randolph Street
Arlington, VA 22203-2114

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
This final report describes the "Leveraging Small-Lexicon Language Models" project, contracted for 18 months under DARPA LORELEI. We focused on Asia-Pacific; a global hotspot of disaster risk with high language density, but few electronic data resources and little off-the-shelf language technology. CRCL provided data and initial analysis for five major families with varied typology: Austroasiatic, Austronesian, Hmong-Mien, Kra-Dai, and Sino-Tibetan (these include about 2,000 languages). We delivered more than 1,000 lects from some 500 distinct ISO 639-3 codes, including over 850,000 lexemes. Data mainly came from smallish, high-quality print lexicons developed for linguistic purposes (language sketch, survey, and comparative analysis); these are the only resources that are widely available throughout the region. Primary effort went to normalizing phonological transcription and semantic glossing (using the MetaForm and MetaGloss frameworks we devised), identifying cognate sets, and producing various types of phonological and semantic analysis of the lexicons; we also distributed a multilingual HA/DR thesaurus of disaster-related terms. All language materials are available for re-use under the CC 4.0 license.

15. SUBJECT TERMS
low-resource language, lexicon, phonological modeling

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
					Doug Cooper
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	SAR	76	**19b. TELEPHONE NUMBER** *(include area code)* n/a

Leveraging Small-Lexicon Language Models

Final Report (CRCL Inc)

Contents

This work was sponsored by the Defense Advanced Research Projects Agency, Information Innovation Office (I2O), Low Resource Languages for Emergent Incidents (LORELEI) program. Issued by DARPA/CMO under contract No HR0011-15-C-01117.

List of Figures

List of Tables

List of Symbols, Abbreviations, and Acronyms

AA Austroasiatic

AN Austronesian

ASJP *Automated Similarity Judgment Project*

CRCL The *Center for Research in Computational Linguistic*, a US 501(c)3 nonprofit organization.

HM Hmong-Mien

ISO 639-3 International standard for *codes for the representation of names of languages – part 3*

KD Kra-Dai

LORELEI The *Low Resource Languages for Emergent Incidents* program.

PanLex *Panlingual Lexion*

PHOIBLE *Phonetics Information Base and Lexico*n

ST Sino-Tibetan

TA Technical Area

WALS *World Atlas of Language Structures*

WN *WordNet*

WPD *World Phonotactic Database*

1. Summary

This document summarizes work completed by CRCL Inc (the *Center for Research in Computational Linguistics*, a US 501(c)3 nonprofit organization) in the period July 1 2015 – December 31 2016 as part of the DARPA LORELEI project, contract number HR0011-15-C-0117. It also provides an overall view of LORELEI, and our role in it.

The LORELEI program intends to advance the state of computational linguistics and human language technology, enabling rapid, low-cost development of capabilities for low-resource languages. These will provide situational awareness based on information from any language, supporting emergent missions such as humanitarian assistance/disaster relief, peacekeeping, or infectious disease response.

LORELEI Technical Area 1 addresses the core research challenge of rapidly developing language processing tools for a language without reliance on large corpora or extensive human annotation efforts. TA1.1 focuses on research and development of novel techniques to discover and use "universal" properties and (typological or other) regularities of languages, reducing reliance on huge quantities of language-specific information for translation, information extraction, or other language technologies. This research area builds on knowledge of the characteristic tendencies and regularities of human language, but is not limited to "absolute" universals that apply to every known language.

As a TA1.1 performer, CRCL's task (as outlined in the *Statement of Work* and listed as a series of deliverable milestones) was to:

- deliver cleaned, normalized, curated lexical data and cognate groupings for 200-250 distinct languages (following ISO 639-3) per year.

We also pursued two general activities on behalf of the program:

- discover and implement means of analyzing and enriching the data sets,
- interact with other performers to help define and enable downstream applications.

These involved enhancing and devising applications for our (and other) small lexicons. CRCL was retained under a one-year contract, and an additional six-month extension. All project results are available for re-use under a Creative Commons 4.0 license.

Problem Description:

The U.S. government does not have hard, language-by-language content data, which might support action or planning, for more than a fraction of the world's 7,000 languages. Existing typological descriptions (e.g. WALS) are sparse, phonological data (e.g. PHOIBLE, with <<25% coverage) is limited, and denotational descriptions (e.g. the single-source Ethnologue) do not include or reference documentary data. This resource gap affects both practical operational concerns – providing actors on the ground with "human intelligence" regarding speaker communities – and long-term strategic technology planning for language-engineering tools: we can't issue a challenge to develop new tools for small-footprint, low-density languages without gold-standard resources to assess their results.

Weighing *cost*, *availability*, and *linguistic value*, the only universally representative, fine-grained resource we might plausibly assemble must be based on the relatively small lexicons (<2,500 words) typically gathered for comparative, survey, or linguistic research purposes. We are assembling and improving this resource for the five linguistic families (totaling about 2,000 languages) that dominate the Asia-Pacific region (28 of the 36 USPACOM countries). This region includes 7 of 10 "global hotspots of disaster risk" (World Risk Report 2013), and has potential for future conflict in restive areas of Myanmar, South China, Northeast India, and insular Southeast Asia.

Expected Impact:

Paradoxically, the best known / most successful languages (e.g. Thai or Vietnamese, for which we have the most resources) are usually poor representatives of the family as a whole. As the only LORELEI project focused on assembling fine-grained language datasets, we contribute to several core problems:

– *identifying training/translation pivot language*s: languages are related to one another by both inheritance – they share a common ancestor, and by contact – one borrows from the other, or both borrow from languages in common. Using the techniques of comparative and historical linguistics and dialectometry will let us suggest which language best represents a given group, and would produce the best results when adapted to a low-density incident language.

– *identifying low-density/small footprint languages*: a low-density language has few computational or analytical resources; a small footprint language is difficult to even find data for. This means that it may be difficult to even identify the language of a potentially important audio or text sample. We provide at least small lexicons for (ideally) half the languages in the region; these may be the only formal resources available for language identification.

– *predicting high-value investment languages*: rather than scrambling to back-fit existing resources to incident languages, we propose that four factors will help predict languages are worth investing in now: a) linguistic centrality, b) currently available resource base, c) speaker population, and d) risk history.

– *producing gold-standard sets of normalized lexical data and cognate assignments*. This provides ground-truth data for future research on rapid development or adaptation of tools and resources for low-density languages.

Research Goals:

Specific goals: the project will extend and apply CRCL technology required to normalize phonological transcription and semantic glossing of a very large number of lexicons (donated to the project in electronic form by CRCL) – and to identify large numbers of related cognate words, which help refine our understanding of (and predictive capacity for) variation between related languages. Our deliverable is the finished product: normalized lexicons and marked cognate sets.

Performance improvements: very few of the world's languages can provide enough electronic data (e.g. via Web pages or social media) to support current computational approaches to language modeling. We will provide the hard data required to produce phonological models, infer etymological and loan relationships, predict word forms (e.g. for entity recognition), and to support unknown language identification.

New capabilities: In narrower terms, the project makes it possible to:

– extract “phonodynamic” language models; that is, phonological and phonotactic sketches whose elements can be weighted against the lexicon for frequency, functional load, salience, phonological neighborhood characteristics, and so on. This is the type of information that helps humans nearly instantly identify even languages they do not speak.

– identify shibboleths; that is, simple words from two or more languages that do *not* resemble each other phonologically, and can be used to help identify speaker language.

– show the linguistic ground path of an expected event; that is, identify the speaker communities that are predicted to be in the path of a typhoon, tsunami, epidemic, or other disaster

– show the human terrain of an ongoing event; that is, identify the speaker communities within the known bounds of an ongoing political or natural crisis.

– build tools for automated orthography-to-phonology; e.g. for generating phonological transcription of L2 dictionaries or texts.

– while project data is at arms’ length from current MT applications, it is reasonable to expect that regular sound-change models will support some named entity identification.

– while project data is at arms’ length from current speech-to-text applications, it is likely to support basic functionality like word boundary recognition.

2. Introduction

Problem description

The U.S. responds to global emergencies of all types. Doing this effectively, safely, and efficiently relies on local, non-English sources of information. But while there are an estimated 7,000 world languages, technology for automated translation, summarization, sentiment assessment and the like is only available for a tiny percentage – perhaps 350 (5%) of them. It is possible to develop such resources language by language, but that is a slow and expensive process, estimated at $10,000,000 each.

Most people speak more than one language; perhaps by choice in the developed world, but as a matter of necessity in the developing world, where one's mother tongue is usually not the language of education and government. . Even though English and the other well-provisioned languages are near-universal *linguae francae* in times of peace (and when people wish to be understood), in times of emergency or conflict (and when people do not necessarily want to be understood) the smaller languages become increasingly important.

Code switching – slipping into a second language in the course of written or spoken discourse – is well-understood not only as a means of concealing information, but as a marker of information that is especially urgent or meaningful. Even if translation technology or a detailed language description is not available, the simple ability to identify any and every language is an important tool. Consider countries like Indonesia, Myanmar, the Philippines, and China (with 700, 117, 200, and 300 languages, respectively). We can readily acquire Twitter feeds and on-line messaging, but messages in, or mixed with, most minority languages will be discarded simply because we cannot classify them.

Surprisingly, perhaps, we do not have digital language models, printed descriptions, or reference samples for most languages. Many websites that purport to provide language documentation on a global scale generally draw from a handful of sources, such as *Ethnologue* or *Wikipedia*, which themselves supply only bare details. The depth of coverage available falls off rapidly, even from sites (*World Atlas of Language Structures*) that are widely cited in the literature. And the type of information provided may be of interest for linguistic purposes, but of little value for computational linguistic applications; e.g. a phonological sketch that contains only a list of phonemes, without any frequency or phonotactic detail.

Traditionally, language technology efforts have worked from the top down, beginning by developing resources and tools for the largest languages (such as English, Chinese, various European languages), and gradually trickling down to smaller languages. LORELEI's predecessor, the REFLEX LCTL project, attempted to extend and accelerate this process (and produced resources used in LORELEI). While the idea of using interlingua *pivot languages* is not new, applications have been limited.

LORELEI attempts to change this equation by recognizing that neither language-specific translation technology nor extensive language resources are necessarily required to obtain *actionable information*. At the extreme, a "peephole" view into communications all that is required to recognize disaster-related words or sentiment. The challenge is not to translate all messages, but rather to recognize high-value messages or messaging.

The *Leveraging Small-Lexicon Language Models* project

CRCL's contribution begins with the broad question *can small lexicons help solve big language problems*? Can minimal, but fine-grained, phonological and lexical data make a useful contribution to both regional and global understanding of language universals and interaction? Is it even possible to develop such data resources so quickly?

CRCL proposed to provide small lexicons for 200-250 distinct Asia-Pacific ISO 639-3 codes per year. Although LORELEI's scope is world-wide, we chose to focus on Asia-Pacific for a variety of reasons, the main one being that it was the largest possible region in terms of languages that could reasonably be managed within the confines of the program.

complexity: extremely high language density	Indonesia: 700, China: 300, Philippines: 200, Malaysia: 146, Nepal 125, Myanmar 117 ...
history: "global hotspots of disaster risk"	7 of 10 highest-risk countries are in Asia-Pacific
risk: likely regions of future conflict	"highland" populations, borders within borders,
responsibility: US Pacific Command region	36 countries, 3,000 languages – **28 / 2000 in our defined area**
infrastructure: a TIPSTER moment	providing linguistic data for the long tail of least-resourced languages

Table 1 Motivations for "Small Lexicon" project design.

The Asia-Pacific region is home to some 3,000 languages: more than 40% of the world's total. Our mission has been to define, and then deliver, the resources that will have the largest impact on language understanding. To do this we have focused on:

- five language families that account for some 2,000 languages, and blanket nearly the entire region from the Himalayas to the South Pacific (excluding Australia and parts of New Guinea),
- small lexicons, typically ranging from 500 – 2,500 words, that were assembled for language survey, sketch, and/or comparative research.

These are typically high-quality resources that provide detailed phonological transcription of all items. We chose to focus on small lexicons for two reasons:

- they are the only nearly universal resource available,
- although they only supply a modest amount of translation, comparative and survey lexicons are the ideal minimal resource for language modeling.

Project details In the context of our project:

- data is almost invariably received in phonological transcription, not formal orthography,
- nearly all data has been previously published (or collected and not published for one reason or another). We are not eliciting new data, or transcribing existing field tapes.
- lists were elicited as part of field or comparative surveys, usually by trained linguists, and are of objectively high quality, especially in contrast to typical "found" data sources, however …
- lists are often *sui generis*, not based on text corpora, or supported by other reference resources; hence, it is not always possible to confirm our interpretation of the authors' intent,
- original lists usually have <2,500 items. Some survey lists will be shorter; many SIL surveys track ~400-500 items, and in some areas 200+ item Swadesh-style lists are all that are available,
- items are usually glossed with a single sense – not defined with multiple senses,
- lemma forms are most common, compounds and complex morphology less so. With a few exceptions, only Austronesian (AN) languages regularly have inflectional morphology; particles and auxiliaries are common in the other families.
- some sources may mark morphological boundaries; these marks are passed through in the raw forms, but not in the normalized forms.

In the 18 months of our project we focused on a relatively small number of sources that provide broad geographic and phylogenetic coverage, and raise a wide sample of typological and notational issues.

Processed CRCL datasets are assembled by running raw inputs through a software system that is frequently tweaked and rebuilt. Datasets are provided both as single aggregated files (one per resource type), and as many files distributed in a *family / ISO / lect* directory hierarchy in XML and TSV formats.

CRCL had three primary tasks in acquiring and working with raw lexicon data:

- add a layer of normalized glosses we call *MetaGlosses*; usually numbered WordNet 3.0 senses.

Metagloss semantics index words that are etymologically related, and whose raw glosses differ only by an authors' choice of vocabulary or phrasing: *rock* versus *stone*. The metagloss will not greatly diverge from the raw gloss even if etymological grouping might call for it. But, in moderately ambiguous situations (*cloudy* versus *gloomy*) we favor the more common term. Most metaglosses are WordNet 3.0 senses [Miller 1995], extended when necessary to fill English-language lexical gaps, or to allow consistent handling of categories like kin terms. On occasion, insight gained from downstream cognate grouping may prompt revision of a sense assignment.

- add a layer of normalized phonological forms we call *MetaForms*.

Metaforms generally make unambiguous substitutions that transform ad hoc notations to (usually) standard IPA notations. As with metaglosses, situations arise in which raw forms must be slightly reinterpreted to achieve the consistency downstream applications require. Normalized metaforms are analyzed into syllables, sub-syllabic components, and individual phonological segments. Our goals are utilitarian, rather than theoretical: to reveal, measure, and if possible extend the form's usefulness for language identification, lexicon extension, cognate identification, audio segmentation or transcription, and similar applications. As with glossing, transcriptions may occasionally be revised with the benefit of information from cognate grouping.

- group etymologically related forms into cognate sets we call *EtySets*.

When possible we seek support and guidance from comparative sets and proto-form reconstructions found in the literature. We do not produce new reconstructions, or attempt to discover long-range etymological relations. We anticipate that the primary use of our cognate sets will be to support applications like lexicon extension, drawing on evidence from predictable, regular phonological variation between relatively closely related languages.

cognate set reference

language metadata

gloss data – silver item is WN 3.0

form data

brief form analysis

detailed form analysis

we show syllable structure and sub-structure, along with positional (phonotactic) information


```
<entry id="hudak2008comparative:C:c11.r151a.g151.i2391">
  <family>KD</family>
  <cogset>KD:H151</cogset>
  <iso>nut</iso>
  <language>Nung (Viet Nam)</language>
  <dialect>Western</dialect>
  <latLong>22.1166,105.5255</latLong>
  <country>Viet Nam</country>
  <adm level="1">Tỉnh Bắc Kạn</adm>
  <gloss status="copper">hammer</gloss>
  <gloss status="silver">hammer#n#2</gloss>
  <form status="copper">hun⁶ thii¹</form>
  <form status="silver">hun⁵⁵ tʰiː¹⁴</form>
  <form status="silver" style="tokenized">:.h.:¦u...¦:.n.:¦⁵⁵ :.tʰ.:¦iː...¦¦¹⁴</form>
  <form status="silver" style="segmented">h.u.n.⁵⁵ tʰ.iː.¹⁴</form>
  <tokens>
    <syllable canon="CV">
      <onset><core>h</core></onset>
      <nucleus><core pos="3.1">u</core></nucleus>
      <coda><core>n</core></coda>
      <tone>⁵⁵</tone>
    </syllable>
    <space />
    <syllable canon="CV">
      <onset><core>tʰ</core></onset>
      <nucleus><core pos="3.1">iː</core></nucleus>
      <tone>¹⁴</tone>
    </syllable>
  </tokens>
</entry>
```


Figure 1 A sample of the information that accompanies each of the 850,000+ delivered lexical items.

A typical result item is shown in Figure **1**. This “easy to use” XML build (from a **lexicon.xml** file) bakes in source and language metadata, shows both raw (“copper”) and normalized (“silver”) versions of the gloss and form, and includes a brief and detailed phonological analysis of the normalized form. This layout can be modified if desired.

This is the type of data required for quantitative and comparative methods of inference of trees of inherited phylogenetic relations, and graphs of loan relations. It lets us address the following kinds of questions (although implementing these was beyond CRCL’s project scope):

- given a basic (200-2,500 words) incident language lexicon in its areal context, can we infer enough details of phonology and morphology to extend functional vocabulary using non-incident language resources?
- given text from a low-resource incident language, can we use a basic lexicon, a language model at least partially obtained from it, and one or more pivot languages to enable translation, named-entity recognition, or other situational understanding?
- can minimal, but fine-grained, phonological and lexical data make a useful contribution to both regional and global understanding of language universals and interaction?

The project raised many other questions and possibilities as well:

- how much information does automatic transcription require?
- how well do wordlists enable phonemic, phonotactic, and morphological language modeling?
- how well does the lexicon reflect an open corpus for these distributions?
- can we anticipate characteristics of difficult-to-obtain corpora; e.g. non-orthographic languages that whose only written appearance is in unmarked informal social media?
- how small a dataset will still produce a useful language model?
- can we devise stopping rules for minimally useful sample sizes? Can we tell when we have enough?
- what types of information can be meaningfully aggregated between small language samples? When can we define clusters of related languages for which this is appropriate?
- how many cognate pairs are required to induce enough parent proto-forms – implicitly, regular rules for sound-change or morphological variation – to accurately remodel existing data?

3. Methods, Assumptions, and Procedures

Data sources and grades

We rely primarily on published materials, although in some cases, linguists will share unpublished texts or data. While born-digital publication and distribution has become more common in the past few years, most sources are traditionally printed (or in the case of some unpublished field notes, handwritten). Nearly all of these resources provide transcribed forms and glosses, and were elicited for language survey, sketch, or comparative research applications. Use of ordinary dictionaries is uncommon.

papers	MKSJ, LTBA, NUSA, JSEALS, OL, PL, other
theses	world-wide, including many Thai, Chinese, other
surveys	may cover closely related lects; e.g. Myanmar
sketches	particularly extensive in Southern China
gray literature	informally published, not widely distributed
field notes	often unpublished / only available source
comparative	Shorto, Blust, Sidwell, Ratliff, Matisoff, Gedney, other
e-resources	MKLP, STEDT, ACD, ABVD
extent	ideally 2,500, but Swadesh if necessary
quality	best available resource, but mileage *will* vary

Table 2 Typical data sources and characteristics.

We use the rough nomenclature in Table 3 to describe data.

vapor	we've heard of it, but haven't seen it
water	untranscribed audio only
paper	paper or pdf, not transcribed or extracted
tin	dictionary e-data: orthography and definitions
copper	comparative / survey e-data: forms and glosses
bronze	some vanilla algorithms naive normalization of forms / glosses, some cognate sets
silver	customized machine processing, machine-usable, but not verified
gold	human-verified, machine-usable, comparable datasets

Table 3 Informal nomenclature used to describe data quality. Our "silver" is in fact linguist-verified and "good as gold" for all practical purposes – we are delaying "gold" assignment until the sets are rolled out to the wider linguistic community.

CRCL brings all copper-standard data to the program: data transcribed as-is, provided in Unicode, with nothing beyond incidental normalization.

3a. Comparative coverage

A number of open-access databases provide linguistic data, but their coverage of the Asia-Pacific region tends to be limited in breadth (few languages are covered) and/or depth (coverage is superficial). This comparison was conducted in May, 2015, and relies on family grouping of ISO codes per Ethnologue 18 [Lewis 2016] (results from Glottolog [Hammarstrom 2016] would be very similar), or the sources' own internally reported grouping (helpful for WALS [Dryer 2013], which does not always map its data to ISO 639-3 codes).

Linguistic Data	ISO 639-3[1]	CRCL Y1	CRCL Y4[2]	WALS (2679)	WALS[3] ≥ 25/10%	PanLex[4] (5963)	PanLex >200	ASJP[5] (4401)	PHOIBLE[6] (2105)	WPD[7]
Austronesian	1257	**109**	626	325	42/160	1060	391	805	42	718
Austroasiatic	170	**30**	85	47	9/23	125	20	93	43	90
Hmong-Mien	38	**19**	19	5	1/3	21	5	15	3	15
Kra-Dai	95	**24**	48	17	3/7	69	9	48	12	33
Sino-Tibetan	474	**108**	242	146	21/87	245	24	165	70	208
Total	**2034**	**290**	**1017**	**540**	**76/280**	**1520**	**449**	**1126**	**170**	**1058**

[1] *ISO item counts are based on the Ethnologue 18 analysis. There are very small inconsistencies in all counts shown because additions, deletions, and modifications to ISO 639-3 are not always migrated to the sources, or because there was uncertainty or disagreement about language identification.*

[2] *Figures in the Y4 column reflect potential CRCL milestone requirements for 40-50% ISO 639-3 coverage. Actual coverage of AA/HM/KD will probably be nearly complete.*

[3] *These figures show depth of coverage. WALS has 194 feature categories; we list the number of WALS datasets that have data for at least 25% and 10% of the WALS feature set.*

[4] *The PanLex [Kamholz 2014] sets in Asia-Pacific are predominantly very small samples (50% have fewer than 45 items). Returned sets appear to be rough synonym sets, and there is no attempt to normalize notation, or differentiate between orthography and phonological transcription. Cited figures in the >200 column count only the largest language variety within any ISO code (these figures are typically inflated by double-counting of the same items from multiple sources; e.g. ASJP and the ASJP source).*

[5] *The ASJP [Bakker 2009] sets contain a maximum of 40 words per lect, written in a reduced phonological transcription. They are also included (and often provide the main data for) the PanLex distribution*

[6] *PHOIBLE [Moran 2014] provides lists of phonological segments with detailed source documentation.*

[7] *The World Phonotactics Database [Donohue 2013] summarizes phonotactic restrictions (e.g. "Is the coda preferentially a nasal?") as +/- binary features, or counts (e.g. "Total vowels"). It does not provide lexical items or transcribed phonological data.*

Table 4 Limited language-family coverage of currently avaliable resources.

A variety of projects and organizations attempt to provide or find ordinary text data for as many languages as possible. It is helpful to bear in mind, however, that the most readily accessible online texts for low-density language are often religious tracts. Like many low-density language Wikipedia pages, they often have a high proportion of transliterated names and toponyms that may skew language modeling unless detected.

The *An Crúbadán* project supplies orthographic trigram models for language identification, as well as word and word bigram frequencies, and links to the discovered text sources [*Scannell 2007*]. It is possible that the paucity of sources for Asia-Pacific texts is due to our inability to properly seed Web crawlers for these texts, or to accurately identify them when they are found.

Corpus data	ISO 639-3	CRCL Y1	CRCL Y4	Scannell (2124) [1]	CRCL Y1 ∩ Scannell	UN (428) [2]	Relig (426) [3]
Austronesian	1257	**109**	626	267 (281)	59	32	116
Austroasiatic	170	**30**	85	14 (14)	2	7	0
Hmong-Mien	38	**19**	19	5 (7)	3	3	0
Kra-Dai	95	**24**	48	6 (8)	5	4	3
Sino-Tibetan	474	**108**	242	67 (72)	23	27	0
Total	**2034**	**290**	**1017**	**359 (382)**	**92**	**73**	**119**

[1] *See the project / download page at http://crubadan.org. The corpus base appears to have been updated most recently in 2015. Figures in parentheses were derived by counting ISO codes on the site. Some of these have been retired, but data appears to have been migrated properly. The next column looks at the intersection between CRCL's Y1 deliverables and Scannell's data (included in our distribution)*

[2] *United Nations Declaration of Human Rights (xml files available at http://unicode.org/udhr/downloads.html)*

[3] *The Watchtower (http://jw.org) has links for 671 lect-specific pages (with fewer distinct ISO codes); we have not finished identifying ISO codes for these. eBible.org (http://ebible.org) links to 545 ISO-specific resource sets. It is likely that the Scannell totals incorporate most of what might be found separately from strictly religious sources.*

Table 5 Text corpus availability for the AA, AN, HM, KD, and ST language families – coverage is about 17.5%.

3b. Metadata

Additional metadata can be associated with each word list. This includes:

- *bibliographic* source metadata: the original text, author, publisher, and other publication details.
- *language* metadata: this includes the ISO 639-3 code and name, an (idealized) speaker location, speaker population, and linguistic subgroup details. Aside from the ISO code and name, all of this information is the result of an independent analysis of some sort. The most authoritative and fully developed analyses have been developed by *Ethnologue* and *Glottolog*; the former is partly open-access and partly licensed, while the latter is open-access. We provide information from both. However, because *Ethnologue* GIS data may not be redistributed, we locate and supply the nearest populated place instead.
- *doculect* metadata: information provided by the author to help identify the published lect; this may include a location, the author's (or speaker's) name for the language, a dialect name, and details about the informant. To the best of our ability we add details about the notation (e.g. *IPA, formal, informal*) and analysis (e.g. *phonemic, broad, phonetic*) used for transcription. Doculect metadata is the basis of the registration of each dataset's *DOI* (digital object identifier).

We take different approaches to providing the metadata: it may be cross-referenced by any dataset that requires it (e.g. used as standoff annotation), or some or all metadata can be baked into each and every set. Please let us know if a custom formulation may be helpful. Figure 2, below, shows a typical metadata set.

3c. Dataset identification and logical tables

For various reasons a single logical lexicon or collection of lexicons may be broken up into separate pieces in a printed work. For example, in some short survey lists each page contains all forms for a single language without glossing. For longer lists, each page may cover only a few words (one per column, with one language per row), or many (with one word per row, with languages labeling columns on one or two pages). And, in some cases, a single set of lists may be split into many tables, as when the author is making a case for a proto-language reconstruction.

We conceive of all the lects in a given text as forming a single logical table when this perspective benefits the user; generally, if they share essentially the same gloss list. In a logical table, lects always

label the columns, and glosses always label the rows, even if the printed work reverses this order. This allows us to uniquely identify each lect with a *bibref* and *column number*, where the bibref is the author's last name, the publication year, and the first non-stop word of the title. The language name appears in the final position for non-English publications, and in cases where a series of similar titles would be confusing.

On occasion, a single text may contain more than one logical table; as when two sets of lects have substantially different gloss lists, present data from different families, or different in the content or presentation of data. In such cases a number is added to the bibref: *bibref_1*, *bibref_2*. Column numbering restarts with 1 in each table. Note that not all columns are necessarily transcribed or provided as part of CRCL's LORELEI data.


```
<dataset id="huffman1971vocabulary.c1">
  <metadata>
    <reference>
      <id>huffman1971vocabulary</id>
      <doi>15144/huffman1971vocabulary</doi>
      <creator>Huffman, Franklin</creator>
      <title>Unpublished vocabulary lists</title>
      <date>1971</date>
      <publisher>Huffman Papers, sealang.net/archives/huffman</publisher>
      <lects>18</lects>
    </reference>
    <language>
      <languageCode scheme="iso639-3">khm</languageCode>
      <languageName scheme="iso639-3">Central Khmer</languageName>
      <latLong source="Ethnologue18">12.4671,104.5699</latLong>
      <latLong source="Glottolog2.6">12.0515,105.015</latLong>
      <country source="Ethnologue18">Cambodia</country>
      <country source="Glottolog2.6">Cambodia</country>
      <adm level="1" source="Ethnologue18">Kampong Chhnang</adm>
      <adm level="1" source="Glottolog2.6">Kampong Cham Province</adm>
      <population source="Ethnologue18">14224500</population>
    </language>
    <doculect>
      <id>huffman1971vocabulary.c1</id>
      <doi>15144/huffman1971vocabulary.c1</doi>
      <creator>CRCL</creator>
      <date>2015</date>
      <notation>IPA</notation>
      <analysis>broad</analysis>
      <forms>887</forms>
    </doculect>
  </metadata>
```


Figure 2: a typical metadata set, showing the bibliographic reference, language, and doculect sections. These may be packaged together with a dataset, or separately as part of a text and data bibliography.

3d. Defective entries

A raw data entry may be excluded from the distribution set for various reasons, including:

- the gloss could not be reliably translated, or there was no reasonable WN 3.0 equivalent or extension available for the gloss (this sometimes occurs for phrasal entries),
- the form could not be reliably normalized or analyzed (this sometimes occurs when the form includes markup or typographical errors).

We can arrange to pass defective entries through if desired.

3e. Morphological information

With rare exceptions, of the five language families we cover only Austronesian has active inflectional morphology. As a rule, the datasets we provide do not regularly mark morphology. Any markup that is provided is explicitly supplied (generally using hyphens, or an occasional parenthesized affixes) in the raw form without further information or analysis.

Some of the Sino-Tibetan data marks apparent etymological affixes. This was usually added to the source data by the STEDT project [Matisoff 2010] in the course of their attempts at reconstruction of proto-Sino-Tibetan. These markers are retained in the raw forms, but should not automatically be understood to be the result of methodical morphological analysis.

In the non-Austronesian families, the use of class terms, particles, phonological and semantic doubling, and other word-compounding processes provides a type of morphology. These will be segregated in due course as we group cognate sets.

3f. Normalization and standardization of glossing

Most of our datasets use glosses to indicate the words used to elicit forms from native speakers, rather than to define and/or explain known native-language words. Frequently, standardized elicitation lists are used. Unfortunately, many glosses, standardized or not, are open to slight reinterpretation by any given linguist or informant. Hence, normalization of glossing is neither trivial nor certain. In most applications, small differences between the gloss, and the item's "true" semantics, will not be critical:

- survey and comparative lists are used to elicit central, core, universal semantic concepts; not subtle distinctions. Hence, the word is not likely to contrast with other semantically linked words in the list; e.g. "stone" as an object versus a material, or "throw" versus "toss" or "fling."
- part-of-speech categories (and the variation in English gloss form they might require) may be determined by context, particularly in non-Austronesian families. We rely on conventional choices, e.g. "blue" and "heavy" are adjectives.
- despite subtle differences from the raw gloss, the normalized gloss reliably aligns with etymologically related items in other word lists, and is able to support downstream applications for cognate identification, distance measurement, lexicon extension, phonological modeling, and so on.

We normalize to WordNet 3.0 senses, because it is a mature, well-developed, and widely used resource, replete with analytical tools, and linked to many other lexical resources. Hierarchical relations, well-defined sense definitions, and corpus-based sense counts also help make WN its own disambiguation tool. Nevertheless, WordNet has gaps. It does not define closed-set vocabulary items, nor does it recognize the regular patterning of some lexical items (in particular, kin terms) that figure heavily in comparative and survey wordlists.

Unavoidably, there are also differences in the way English and other languages lexicalize concepts, actions, or things; e.g. "hand/arm" and "blue/green" are indivisible lexical items in much of Asia-Pacific. And, in some cases, we are not sure whether or not a lexical gap exists. For example, "big basket" might be a noun with modifier, a single lexical item distinct from a small basket, or just the standard word used for baskets (i.e. the elicitation list might request "big basket" and "small basket" and receive the same form for both).

Our *MetaGloss* system addresses these issues.

- when possible, a single WordNet 3.0 sense is provided: **house#n#1**
- when two or more useful interpretations are plausible, they are pipe-separated: **bake#v#1|toast#v#1.**
- several word classes have been added (with all items numbered #1): **d**(*emonstrative*), **j**(*conjunction*), **k**(*in term*), **m**(*odal*), **p**(*ronoun*), **q**(*interrogative*), **x** (*temporarily uncategorized*).
- when new senses are added to the WN **a, n, r, v** lists, they are numbered #0: **armspan#n#0**.
- a polysemous sense that does not exist in English is indicated by labeling the WN 3.0 sense: **v@fist#n#1** indicates the verb sense of the noun "fist," i.e. "make a fist."

- kin terms are built up in regular fashion, starting with the person who is ultimately referenced: **mot.fat#k#1** is the mother of the father, or the paternal grandmother.
- senses may have *attributes* that help document what we believe is the useful reference meaning; e.g. **carry#v#1:tumpline**. This indicates that for purposes of cognate grouping the item clusters with "carry" terms, but keeps "tumpline" accessible. These head+attribute forms may be simplified in the future.
- classifiers are noted by the *:clf* attribute, e.g. **basket#n#1:clf** is a classifier for baskets, **several#a#1:clf** for several items, **kick#v#1:clf** is an instance of kicking. There may be some inconsistency in the listing of feature-oriented classifiers (e.g. long, thin items) because it is not always clear if the given form is a classifier, or just an instance of an item.

All senses used in any distribution may be found in the top-level `metagloss/` directory.

3g. Normalization and analysis of forms

There is an enormous amount of variation in the way that phonological forms – even for the same items – are transcribed in the source data. This is due to differences in:

- **analysis** a *phonetic* transcription most closely follows actual utterances. An analyzed *phonemic* transcription ignores allophonic variation and produces somewhat idealized forms. A *broad phonemic* transcription ignores obvious minor variations, but does not guarantee a minimal phoneme set. It is not always possible to ascertain which analysis a transcription relies on.
- **notation** an *IPA* transcription follows the formal IPA guidelines (and directly maps to Unicode glyphs), with some rare exceptions and national variants. A *formal* transcription may pre-date modern IPA practice; it can usually be mapped to modern IPA. An ad hoc *informal* transcription typically uses the roman alphabet, but does not always follow any recognized conventions.
- **tradition** the IPA provides notation, but does not define its usage. Some linguists will suppress features they feel are predictable within the language, while others mark them explicitly. It is not always possible to determine which path has been followed.

CRCL's *MetaForm* normalization has a dual goal:

- to make data *comparable*, despite have been originally prepared using different analyses, notations, and traditions,
- to add an explicit *analysis*, often based on our knowledge of the individual language, that will benefit downstream applications such as cognate alignment, language distance measure, and audio segmentation.

We accomplish this dual goal by:

- **normalization:** translation into appropriate IPA notation,
- **syllabification:** marking of syllable boundaries, which is often needed for proper segmentation,
- **sub-syllabification:** marking of onset, nucleus, and coda syllable segments,
- **segmentation:** division into individual phonological segments – logical single-character entities that cannot always be represented in IPA / Unicode,
- **feature analysis**: specification of the phonological features of each segment, and
- **role analysis:** specification of the position / phonotactic role of each segment.

For example, the imaginary raw form /*mboa*/ may actually vary in length from one (/^{m}*boa*/) to three (/*m*^{a} *bo a*/) syllables. The leading /m/ might be *prevocalized* (/*em*/), *unvocalized* (/^{m}/), or *vocalized* (/*m*^{a}/), according to implied phonotactic restrictions. Similarly the language might allow or forbid diphthongs. MetaForm makes any analysis we are able to provide explicit.

Four characters – / ɿ ʮ ʅ ʯ / – that are not strictly IPA (but which could be replaced by IPA sequences) are retained because they are widely used in the region's modern notation. In effect, they fill gaps that, arguably, the IPA could have provided. One additional character – / *ʋ* / – is used as the high, back, rounded, fricated vowel. It appears variously in the literature as /*v*/ with an over/under diacritic (e.g. /*v̩*/), and there is no formal (or ideal, albeit informal) IPA alternative (e.g. / u^{β} / β̩ / β̞ /).

Syllable boundaries cannot always be determined. In some cases linguists disagree, and in others we do not have the information required to recognize that, for example, a /*-tt-*/ sequence should be a geminate /*-tː*/ rather than /*-t t-/*. To help minimize the consequences of an incorrect choice, we provide all items both in fully tokenized form, and in a simpler rendering as phonological segments. From an earlier example:

```
<form status="silver" style="tokenized">:.h.:¦u...¦:.n.:¦55
:.tʰ.:¦iː...¦¦14</form>
<form status="silver" style="segmented">h.u.n.55 tʰ.iː.14</form>
```

The *tokenized* form is easily rendered as sub-syllablic ngrams, while the *segmented* form is trivially converted into ngrams of phonological segments or features.

3h. Feature analysis

CRCL's feature analysis is shown in the Appendix, and partly summarized below. This table drives all feature assignments, and is designed for clarity in tagging tokens, and convenience in downstream applications. It does not account for all possible linguistic behavior worldwide, but intentionally limiting its scope to features characteristic of our five language families of interest helps reveal errors in data input or analysis: they require impossible tokenization or feature assignments. All token-to-feature assignments are unambiguous and reversable. Note that some phonotactic information (e.g. **role** and **position**) is built in.

Category	Attributes
class	*consonant, vowel, syllabic, minor*
role	*onset, nucleus, coda*
position	*core, post*
length	*epenthetic, short, long*
pre-articulation	*prenasalized, devoiced, preglottalized, preaspirated, prelabialized, prestopped*
height	*high, near-high, close-mid, mid, open-mid, near-low, low*
backness	*front, near-front, central, near-back, back*
place	*bilabial, labiodental, dental, alveolar, retroflex, palatoalveolar, alveolopalatal, palatal, labiopalatal, velar, labiovelar, uvular, pharyngeal, glottal*
manner	*nasal, stop, implosive, affricate, fricative, approximant, tap-flap, trill*
realization	*rounded, voiced, retroflexed, lateralized, fricated, nonvocalized, prevocalized, vocalized*
phonation	*nasal, aspirated, devoiced, breathy, creaky, dental, raised, lowered, rhotic*
post-articulation	*nasalized, glottalized, palatalized, labialized, labiopalatalized, stopped, velarized, pharyngealized*

Table 6 Main features of CRCL's phonological feature analysis. This is provided in full in the Appendix.

The **class** attributes *syllabic* and *minor*, and their associated **realization** features *nonvocalized*, *prevocalized*, and *vocalized*, are specifically intended to address the problem of inconsistent notation of unstressed onset syllables (*sesquisyllables*) widely found throughout the region, e.g. /*kka*/, /*ka ka*/, /*k ka*/,

/*k̩ ka*/, /*k.ka*/, /*k*^{ə} *ka*/. As a rule, when onsets clearly violate the sonority sequence principle, we treat them as minor syllables, without overt vowels, whose vocalization might or might not be inferable from our knowledge of the language and/or the author's transcription practice.

This has a number of advantages, not the least of which is simplifying automated cognate segment alignment and distance measurement. One consequence – which we accept, because it is characteristic of all families that we work with – is that complex onsets that violate sonority are not seen. We accept this with the understanding that this analysis may be extended in other areas of the world.

3i. Phonodynamic inventories and ngrams

Phonodynamic analysis datasets supply lect-by-lect surveys of phonological segments, their positions within syllables and words, and various statistical measures. They allow the inference of phonotactic restrictions on (or preferences for) segment collocations. However, it is important to understand that these are purely data-driven. They should inform, rather than substitute for, a formal analysis.

We supply two basic phonodynamic dataset types; one of tokens, and one of features. For the moment, they are both in TSV (not XML) form. Below, a token survey (a similar table laid out by rows is also provided) that shows:

- counts for sub-syllable tokens: the complete nucleus, onset (**onCC**), coda **(codCC**), and tone contour),
- counts for individual segments, by position (for consonants) or value (for vowels),
- summary counts of each syllable canon.

```
hudak2008comparative 1  tha      Thai
nucleus onCC     codCC    vow      core     post     onCore   onPost   codCore codPost canon      tone
a 187   kl 23    -        a 309    b 28     l 52     b 28     l 52     j 122   -       CCVCT 60   ²² 273
aː 222  kr 2     -        aː 226   c 18     r 33     c 18     r 33     k 105   -       CCVT 7     ²⁴ 145
e 38    kʰl 5    -        ă 17     cʰ 24    -        cʰ 24    -        m 84    -       CCVVCT 9   ³³ 258
...
```

Figure 3 Counts from *sketch-cols.tsv*. This provides a quick overview of phonological and sub-syllabic segments.

The second basic type provides a segment-by-segment feature inventory, also with positional counts.

- counts for each token, by position: 1-4 for vowels, or onset, coda, or minor syllable onset or coda,
- a tabulation of each segment's phonological features: length, pre-articulation (e.g. pre-nasalization), height, back, place, manner, realization (e.g. rounding, voicing), phonation (aspiration, creak, etc.), and post-articulation (e.g. palatalized or glottalized).
- summary counts of all n-thongs, onsets, codas, tones, and syllable canons are also provided.

```
hudak2008comparative 1  tha     Thai
Token total 1/onset 2/coda  3/minOn 4/minCo length  pre-art height back    place   manner realize phonat
post-art
a     309   193     116                                  low    central
aː    226   226                          long            low    central
ă     17    17                           short           low    central
      ...
b     28    28                                                        bilabial stop  voiced
c     18    18                                                        palatal  stop
cʰ    24    24                                                        palatal  stop         aspirated
d     51    51                                                        alveolar stop  voiced
      ...
N-thong total
ia       31
ua       35
ɯa       48
CC onset total
kl       23
kr       2
kʰl      5
      ...
```

Figure 4 Counts from *sketch-features.tsv*. This provides an overview of segment features by position, and multi-segment onset, nucleus, and coda sections.

Many statistical measures of feature significance are calculated. Because these are based on simple calculations using unweighted samples, they must be viewed as extremely rough indicators. They include:

- **diphone/triphone frequency vectors**: their orthographic equivalents are very effective for text language identification; it is not clear if wordlist distributions are enough to characterize language similarity. We generate these for both segments and specific features (e.g. consonant **place** and vowel **back** collocations).
- **functional load:** a measure of the segment's information content; how necessary is it to uniquely identify its context? We calculate this as the segment's number of *contrastive / total* appearances; i.e. the number of times that the segment must be known to disambiguate a lexeme divided by its total appearance count. (See also [Surendran 2003, 2006].)
- **salience**: the equivalent of *inverse document frequency* [Sparck-Jones 1972]; how well does a particular segment or collocation identify a language? By treating each language's list of segments as a document, we can define each document collection as the set of languages within a given geographical (i.e. *n*00-mile radius) or etymological (e.g. sub-branch sisters) distance from the target language. Thus, salient segments may provide geographic shibboleths, or evidence of shared etymological innovation or loans.
- **neighborhood and clustering coefficient**: how closely linked (i.e. varying from one another by a single feature or segment) are the words in a language, and what is each word's phonological *neighborhood*? [Vitevitch 2007, Luce 1998] Because we expect sound changes to be regular, we expect neighborhoods to be recognizable even if surface forms vary. Thus, this data can serve as a proxy for language divergence.
- **wordlikeness**: how well does a word reflect both the phonological distributions and phonotactic constraints of a given language?

We have extracted a series of unigram and ngram sets from the data, by lect. These include:

- phonological segment bi- and trigrams: implicit blanks before and after each word are treated as segments. (**`2_segment.tsv, 3_segment.tsv`**)
- segment(s) plus nucleus bi- and trigrams: these treat the nucleus as a single phonological segment. (**`2_segment_nuc.tsv, 3_segment_nuc.tsv`**)

- sub-syllabic (onset / nucleus / coda and coda / onset) bi- and trigrams: again, implicit pre- and post-syllable blanks are treated as tokens. (**`2_token.tsv, 3_token.tsv`**)
- onset or nucleus plus tone collocations: these are only calculated for tone languages. (**`2_onset_tone.tsv, 2_nucleus_tone.tsv`**)
- feature trigrams: these separately track (consonant) place and (vowel) backness, and (consonant) manner and (vowel) height. (**`3_place_back.tsv, 3_manner_height.tsv`**)
- functional load, by phonological segment: these count appearances and contrasts, and calculate load (**`load.tsv`**).

Other ngrams can be extracted on request.

3j. Lexical analytics: contrast, cover, neighbor, wordlikeness

Lexical analytics describe the relationship between forms, and between forms and the full lexicon. We have extracted min contrast and min cover sets for each doculect:

- minimal *contrast* sets are items that differ by single phonological segment pairs, and are useful for establishing formal phonemic analyses; i.e. recognizing allophonic variation. We list these by segment pair, including the null (e.g. *ball, all*) segment. (**`contrast.txt`**)
- minimum *cover* sets are lists of words that, together, include all segments. These are not unique; more than one possible list may include all segments. This is a computationally expensive operation; we employ a greedy algorithm that is almost certain to return the shortest possible list. (**`cover.txt`**)
- *neighbor* sets treat each word as the central node in a graph; each edge represents a distance of one phonological segment. We calculate the neighborhood density, number of edges, and *clustering coefficient* (number of links between the neighbors). (**`density.tsv`**)
- *wordlikeness* indicates how well a word matches the phonological distributions and phonotactic restrictions of the lexicon as a whole. Although more typically used to evaluate pseudowords, this measure can assist language identification. (**`wordlike.tsv`**)

3k. Related text data

When available, we have included corresponding data from Scannell’s *An Crúbadán* project;:

- trigram grapheme lists, including implicit onset and follower spaces,
- monogram and bigram wordlists,
- source URLs (Scannell does not release the original texts, but provides the links needed to scrape them).

These sets have several applications:

- language subgrouping: distance measures between ngrams (e.g. cosine distance) can be used to generate trees of language relations.
- ortho-to-phono and vice versa: the phonological sets can help build conversion tools when used in conjunction with orthographic ngrams, Among other applications, these will help answer the question of just how well the lexicon reflects the language as seen in a text corpus.
- language identification: it is an open question whether ngrams encapsulate the same kind of phonotactic information that humans rely on for rapid language identification.

We very much want to extend available text data beyond those sets trivially identified by BCP-47 style script codes, or found in Wikipedia pages; see **Web corpus acquisition** in the **Applications** section.

3l. Cognate sets

Cognate sets are provided as standalone XML entries (**figure 5**). All cognate relations are tabulated in **`cognates/grid.tsv`** , which is essentially a table whose rows are ISO 639-3 codes, and whose columns are rough historical glosses, given as WordNet senses. Sets of corresponding items from two or more languages are suitable as training data for applications like inference of regular sound change correspondences, and lexicon extension.

```
<cognate id="huffman1971vocabulary:C:c13.r625.gs2041.i8527" iso639-3="lbo"
lang="Laven">
   <etygloss>roast#v#1</etygloss>
   <etyset>AA:S2041</etyset>
   <form>buh</form>
</cognate>
```

Figure 5 A typical cognate entry. The id provides a unique link to a data item. Language-related details are baked in for convenience, and can be extended if desirable.

The **`<etygloss>`** element provides a nominal index term for all of the cognate clusters with the same rough semantics. This is a term of convenience, and might not actually reflect the meaning of the proto-form. The **`<etyset>`** element identifies the proto-form's nominal family source (here, Austroasiatic), and numbers the cognate cluster. When possible, the number refers to an established cognate set from the literature. Here, **S2041** refers to Shorto's set 2041. Our current reference set includes:

- AA Austroasiatic [Shorto 2006]
- **AN** Austronesian [Blust 2010, Wolff 2010, Greenhill 2008]
- **HM** Hmong-Mien [Ratliff 2010]
- **KD** Kra-Dai [Hudak 2008, Pittayaporn 2009, Weera 2000, Norquest 2007]
- **ST** Sino-Tibetan [Matisoff 2010]

Many cognate sets also have ad hoc identification numbers (e.g. **AA:4**). Items in these sets form a coherent group that is either not reported in the literature (which is hardly exhaustive), or which will probably be moved to a different etygloss set. We derive cognate sets in the following manner:

- calculate the surface similarity between all forms with closely related semantics. We use Kondrak-style phonological similarity, which is robust in the face of feature (vs. IPA character) variation [Kondrak 2002],
- use different clustering algorithms (bottom-up agglomeration, and Markov chain clustering [van Dongen 2000]) to form likely cognate groups. It is difficult to predict what algorithm and parameters will create the most realistic clusters; we pre-calculate a half-dozen trial settings, then choose a starting set,
- individually revise the automatically generated groups, adding references to sets established in the literature when possible.

Many cognate sets will be relatively small at first. We may not yet have data from other languages in the same etymological subgroup, might not have established enough clusters to support claims regarding more dramatic phonological changes, and/or have not yet established a large enough number of sets to reliably merge groups that require an argument for semantic shift.

Formal cognate relations are not always needed to compare wordlists from sister languages that are known to be etymologically close, particularly if they have been elicited using the same glosses. Anybody can perform the same item-by-item distance measure, using their own cutoff rule of thumb for assumed cognate status. However, this simple approach becomes progressively less reliable as the distance between languages increases, or as individual linguists' practice in data collection varies.

Finally, we mention in passing that formal Swadesh lists are not intended to elicit cognates, but rather to expose the rate of cognate replacement. Nevertheless, some comparative surveys may use Swadish or

similar elicitation terms to seek cognates only. Each approach addresses different goals; our point is simply that one should avoid making assumptions about list content and utility.

3m. HA/DR thesaurus

The Ariel project's *HA/DR Topic Lexicon* lists roughly 34,000 terms "*relevant to the HA/DR topic taxonomy devised by DARPA and the LORELEI evaluation team.*" We have extracted a thesaurus of terms that appear both in this list, and as CRCL metaglosses.

We have further extended the HA/DR list by 200+ terms which appear in our wordlists and appear to be relevant, including *kill, poison, nauseous, afraid, fear, grave, blood, bury, hungry, thirsty,* etc. These all have high negative scores in the *SentiWordNet, SentiWord*, and/or *Valence, Arousal, Dominance* analyses [*Gatti 2013, Baccianella 2010, Warriner 2013*]. We think these terms are more likely to be relevant in monitoring informal communications such as Twitter.

4. Results and Discussion

Overview

Datasets provided for the final milestone are summarized in Figure 6, and go well beyond the contract requirements.

Overview of LORELEI data

Language family summary

Family	ISOs	Sets	Cogs	Forms	Total ISO	Coverage
AA	30	50	209	42633	170	17%
AN	335	680	438	550142	1257	26%
HM	19	34	458	14449	38	50%
KD	23	54	249	71548	95	24%
S					14	0%
ST	108	306	375	194616	460	23%
Total	515	1124	n/a	873388	2034	25%

ISO counts are unique within each family (not source).
***Cogs** gives EtySet (concept) counts; each usually contains several distinct cognate groups.*

Source summary

Family	Source collection*	ISOs	Sets	Forms	Avg**	Gloss status	Form status	Notation	Analysis
AA	huffman1971vocabulary	16	18	11997	666	silver	silver	IPA	broad
AA	huffman1979vocabulary	7	11	15481	1407	silver	silver	IPA	phonemic
AA	theraphan2001languages_1	10	14	9420	672	silver	silver	IPA	phonemic
AA	theraphan2001languages_2	6	7	5735	819	silver	silver	IPA	phonemic
AN	arnaud1997lexique	34	36	33186	921	silver	silver	formal	broad
AN	reid1971philippine	40	43	17359	403	silver	silver	IPA	phonemic
AN	reid2016philippine	49	79	33622	425	silver	silver	IPA	phonemic
AN	stokhof1980holle	153	280	244215	872	silver	silver	adhoc	narrow
AN	tadmor2015jakarta	15	52	85925	1652	silver	silver	IPA	phonemic
AN	tadmor2015languages	12	30	15024	500	silver	silver	IPA	phonemic
AN	tryon1995comparative	80	80	90438	1130	silver	silver	formal	broad
AN	yap1977comparative	73	80	30373	379	silver	silver	IPA	phonemic
HM	ratliff2010language	11	11	4782	434	silver	silver	IPA	phonemic
HM	wang1995miao	18	23	9667	420	silver	silver	formal	phonemic
KD	hudak2008comparative	12	18	14163	786	silver	silver	formal	phonemic
KD	zhang1999zhuang	14	36	57385	1594	silver	silver	IPA	phonemic
ST	huang1992tbl	45	49	82632	1686	silver	silver	formal	broad
ST	lsm2015chin	23	142	59566	419	silver	silver	IPA	narrow
ST	lsm2015naga	14	81	29081	359	silver	silver	IPA	narrow
ST	marrison1967classification	28	34	23337	686	silver	silver	adhoc	narrow
	20 source files	660	1124	873388	777				

ISO counts are unique within each source (not family). Minimum count for inclusion is 100 items
**Some source collections may contain multiple bibliographic sources (bibrefs).*
***Roughly equals the number of distinct glosses per elicitation set*

Figure 6 Overview of final deliverable set. As noted earlier, both glosses and forms are gold-standard in all but name – we feel that a formal roll-out, and comment period in the linguistics community, is appropriate.

An overview of the delivery hierarchy is given in Figure 7. The project's data delivery formats evolved rapidly in order to better expose the content of the data sets. Extracting data was not the issue; rather, it was helpful to clarify the different views and data subsets that *could* be extracted.

MetaGloss and MetaForm

There were few surprises in regard to the planned work of the project. We set an extremely challenging schedule, on average processing one ISO code per day, often with two or more lects per code. Normalizing to the *MetaGloss* and *MetaForm* frameworks required a massive amount of effort simply because even with experience and computational assistance, delivering > 850,000 items put us at the wrong end of the lever. Even very low problem rates produced many, many thousands of items requiring individual attention (and sometimes revealing errors in the original data source).

The difficulty of defining a "final" *MetaGloss* standard came as something of a surprise. While it is possible to restrict the content of elicitation sets (such as Swadesh, various regional SIL survey sets, the

```
crcl/ – root directory
   ./formats – description of all document formats
   ./paths – grep-able list of paths to all files
   ./tokens.xml, ./tokens.tsv – all lexical data
   ./sketch-rows.tsv, ./sketch-cols.tsv, ./sketch-features.tsv – all segment/canon/feature overviews
   ./readme.panlex – notes on and aggregated manifests for Panlex data
   bib/ – bibliographic metadata
      ./metadata.xml
   geo/ - geographically oriented data
      ./info.geo – list of family, ISO-639-3, county, and ADM-1 region (if available)
      CN/ - one directory per country, ISO 3166-1 alpha-2 codes
      KH/ – ... (about 25 countries in all)
         ./info.geo – country summary (ADM-1 regions are not always available)
         ./Champasak.geo – one file per ADM-1 region.  These may later be changed to ISO 3166-2 codes.
         ./Preah_Vihear.geo ...   etc.
   metagloss/ – global data for MetaGloss (WordNet 3.0 glosses)
      ./metagloss.txt – all forms and counts in use
      ./new.txt – list of new (sense 0) items
      ./kin.txt – explanation of the components of kin terms
      ./a.txt ... x.txt – lists, by part of speech, for all items
   cognates/
      ./cognates.xml – single file of all items with tagged etygloss and etyset
      ./setByRow.tsv – training data table of all cognate relations (columns are lects)
      ./setByCol.tsv – training data table of all cognate relations (rows are lects)
      etygloss/
         able#a#1/ – one directory per concept/label. 200+ sets per family Y1 to 500 Y4
         above#r#2/ ...  Not all sets overlap, and we substantially overshoot the targets.
           ./etyset-1.xml – one file per etymologically related set; typically several
           ./etyset-n.xml ... sets per concept, per family; e.g. AA:S638.xml, HM:R837.xml
   hadr/ – extended HA/DR-specific lexicon, across all languages
      ./readme.txt – discussion of HA/DR item acquisition and form.
      crcl/, panlex/ – one directory for each major source
         ./readme.txt – source-specific notes
         ./hadr.tsv – comparable lexicon
   AA/– one directory each for Austroasiatic, Austronesian, Hmong-Mien, Kra-Dai, Sino-Tibetan
   AN/, HM/, KD/, ST/ ...
      alk/ – one directory for each 3-letter ISO 639-3 code; expect 250+ Y1 to 800-1,000++ Y4
      brb/ ...
         arnaud1997lexique.c1/ – one directory for each documented lect, where directories
         arnaud1997lexique.c2/ ... are named as bibref.column.  500 doculects Y1 to 2000 doculects Y4
           ./metadata.xml – metadata for this lect
           ./lexicon.xml – main lexicon file
           ./sketch-cols.tsv – sketch of segments, column view (easier to read)
           ./sketch-rows.tsv – sketch of segments, row view (easier to grep)
           ./features.tsv – sketch of segments by their features
           ./2_segment.tsv, ./3_segment.tsv – phonological segment bi- and trigrams
           ./2_segment_nuc.tsv, ./3_segment_nuc.tsv – phonological segments, single nucleus
           ./2_token.tsv, ./3_token.tsv – sub-syllable token trigrams (onset, nucleus, coda, tone)
           ./3_place_back.tsv – place/back feature trigrams
           ./3_manner_height.tsv – manner/height feature trigrams
           ./2_onset_tone.tsv, 2_nucleus_tone.tsv – onset / nucleus plus tone collocations
           ./cover.tsv – minimum cover set
           ./contrast.tsv – minimal contrast set
           ./density.tsv – clustering coefficient, links, degree, neighbors for each word
           ./load.tsv – functional load, by segment
         info/ – other language data relevant to the ISO 639-3 code
           ./metadata.xml – metadata from Ethnologue, Glottolog.
           ASJP/  – one directory for each wide-coverage source
           Ethnologue/ ... this anticipates we may rely on or develop other sources
           Glottolog/  ... a typical example:
             ./geo_distance.tsv – geographical distance sets (0 to 500 km, by 100km)
             ./ety_distance.tsv – genetic distance sets (n nearest neighbors)
             ./geo_lexicon.tsv – lexicon of all neighbors within 250 km; known cognates marked
             ./ety_lexicon.tsv – lexicon of all of this ISO code’s sisters
           Panlex/
             ./manifest.tsv – summary listing of count, source, quality, license for all lect data
             ./iso-var.tsv – PanLex designation of the lect, e.g. tha-001.tsv
         text/ – orthographic data if available
           Scannell/ – at present, only files from the An Crúbadán project are supplied.
             BCP-47/ – the sample’s BPC-47 code
               ./info.txt – lect and source data identification
               ./urls.txt – sources for the ngrams and wordlist (texts are not included)
               ./chartrigrams.txt, ./wordbigrams.txt, ./words.txt – datasets
```

Figure 7 Structure of the distribution. When appropriate, files have a comment that recapitulates source information, so that full sets can be concatenated from the root, e.g.: CRCL/% **cat `find ./crcl | grep 3_segment.tsv` > 3_segment.tsv**

IDS / LWT family, and the ILCAA / Princeton family), we faced the opposite problem of having to accommodate a wide variety of formal and informal gloss lists. We see *MetaGloss* remaining as a restricted but extensible framework rather than a completely controlled standard.

The *MetaForm* feature analysis, in contrast, converged fairly quickly on the set now in use. Nevertheless, we had to retain some notational features (the "Chinese" IPA characters) whose importance might not have been obvious had we begun work in a different region. Thus, we anticipate that, say, the African languages will call for both predictable and perhaps unpredictable extensions.

Process management

Finally, we noticed an interesting degree of culture clash between computational and comparative linguists, both within our team, and the LORELEI project at large.

computational linguists	***(mostly comparative) linguists***
big data – need for large samples	small data – need for high accuracy
noise that could be ignored	mistakes that needed to be fixed
orthography, reliance on source as-is	phonology, need to modify the given forms
data-driven methods	analytical methods
anonymous discovery / acquisition of data	personal relationships with linguists
difficulty recognizing GIGO situations	desire to build Swiss watches
acceptance of continuous revision	focus on final publication
if it's measurable, it's progress	question if small improvements will scale up
iterative process – rebuild the data system	linear process – assemble final components
linguists should enable better software	software should enable better linguists

Table 7 Typical gaps in perception between computational and comparative linguists.

Our work – methodical selection and normalization of representative data sets – is typically the domain of comparative linguistics and proto-language reconstruction; traditionally an area of boutique / handicraft linguistics. We were interested in finding ways to industrialize this; not simply by building faster software whose output would require less correction, but by providing faster, more accurate data management by the linguists – less "linguists enable software," and more "software enables linguists."

For example, choices made in normalizing notation affected automated syllabification; while tweaks of language and subbranch-specific syllable-break rubrics affected proper recognition of sub-syllabic segments – which sometimes required going back to the beginning and altering notation. Similarly, source glosses were sometimes ambiguous in ways that could only be resolved at the end of the process, when items were being clustered into cognate sets; again, initial source data (glossing) was somewhat indeterminate until the end of the process.

Thus, instead of focusing on standalone software systems that would incorporate linguistic knowledge per se (the "linguists enable software" approach), we also wrote tools that provided myriad data views to expose different kinds of inconsistency, and let the linguist manage the development cycle very, very quickly; e.g. by immediately seeing the ultimate effects of early choices in data preparation, and by fixing the software process, rather than fiddling with the end of the data pipeline. Providing rapid feedback loops on the data life cycle, and constant willingness to redesign tools as needed, made the difference.

5. Conclusions

This document summarizes work carried out by CRCL on behalf of the DARPA LORELEI project. We have described both the specific contract deliverables and our additional activities. All required milestones were surpassed, and all data and analysis is available for re-use.

While the project was limited to providing data for a single region, we have shown that it is possible to develop large-scale, fine-grained, comparable lexical and phonological data sets quickly, and at a reasonable cost. In addition, we have demonstrated that such data has downstream applications in supporting DARPA's mission. We feel that an ongoing project of this type for Asia-Pacific and other regions is both feasible and desirable.

Our present language technology situation hardly seems tenable: for the majority of world languages, we have little data beyond ISO 639-3 identifiers, brief prose descriptions, and rough speaker areas (unfortunately, not defined in terms of standard ADM area boundaries). Specific language data that would be useful in computational applications – dictionaries, grammars, phonotactic analyses, corpora – is only narrowly available.

Experience shows that neither the marketplace nor traditional scientific funding agencies are likely to fill this gap. From the commercial point of view, small languages do not justify investment costs; their speakers are either too few in number, or too poor, even when they number in the millions. From the research point of view (e.g. the NSF-NEH *Documenting Endangered Languages* initiative), funding tends to support documentation of single languages, and the opportunity this provides for training young linguists. When broader linguistic surveys are done, they usually focus on data of phylogenetic interest for proto-language reconstructions that involve single subgroups or families – not on on-the-ground reality that is needed for computationally useful modeling.

To paraphrase Chamfort,[1] we may begin by choosing the most inviting languages, but in the end we want them all. LORELEI is one of a continuing series of exercises in developing language technology. Methods and goals have changed in the decades since TIPSTER, but the list of languages of interest always gets longer.

[1] "Most compilers of anthologies of poetry or epigrams are like people eating cherries or oysters: they start by picking out the best, and end up eating the lot." Nicolas-Sebastien Chamfort, *Reflections on Life, Love and Society* (1795).

6. Recommendations

We conclude with recommendations for ongoing work (beyond extending language coverage).

language identification language identification based on trained trigram models or similar is extremely effective; see [Scannell 2007]. However, we may not have substantial, identified text samples to work with; e.g. when the use of informal orthographies for online / text message communication is widespread, as is increasingly the case for non-roman scripts, as well as languages without formal writing systems. It would be useful to see if a *phonodynamic* language model, based partly on recognizable segments, and partly on the relations, co-occurrence restrictions between, frequency, salience, and functional load of arbitrary segments, is sufficient to identify a language that relies on an unknown orthography.

Web corpus acquisition building text corpora by Web crawling and scraping is a well-established discipline. However, it does not address the problem of crawling and language identification absent a set of seed search terms. Nor may these be trivially obtained if and when a language either has no formal writing system, or is so obscure that, say, its Wikipedia page does not point to native-language sources. We propose that informal low-density language texts are likely to be written using the roman alphabet, and that we can make reasonable guesses as to how our phonologically transcribed data might be transliterated by native-language speakers, providing the necessary seed search terms.

ISO 639-3 audit this standard was adopted in 2007, based on the then-current edition of Ethnologue. It is managed as a completely separate entity, and relies on outside requests for additions, deletions, and other changes. ISO 639-3 does not document languages per se; it points to outside authorities (at this point, only Ethologue) for assistance in *language denotation*, i.e. any descriptive information about the language, or its place among related languages. Ethnologue, in turn, does not regularly document the sources of its conclusions (and has recently gone to a fee-for-access model for these).

The problematic bottom line is that there is no clear measure of the distinction between assigned ISO codes (languages that are essentially the same may have the same code), or of the tolerable degrees of divergence with a single assigned ISO code (so-called dialects may be mutually unintelligible). Government decisions that rely on ISO codes as a measure of linguistic diversity may not be well-founded. CRCL wordlists – in some cases, representing many lects within a single “language” – can show the degree of lexical diversity (or lack thereof) between lects and languages, and lay the foundation for more reliable measures of linguistic divergence.

lexical item generation approaches to this problem include: straightforward machine translation (phonological segments are treated as words in a sentence), extended MT approaches (e.g. adding feature bundle information), or translation by phonological transliteration/transduction. Linguistically motivated approaches include attempting to generate a parent proto-form first (then using that as the translation/transliteration source), and working from an existing proto-language model.

identification or prediction of nativized loanwords while similar to the problem above, this requires a separate analysis that attempts to model the phonological reduction or feature insertion typically found in loanword acquisition (as opposed to the regular, lexicon-wide patterns of phonological variation found in divergent languages).

ortho-to-phono CRCL wordlists provide the necessary data for alignment with dictionary headwords, based on a combination of (raw and normalized) gloss/definition and unambiguous IPA/orthographic correspondences. This should be sufficient for training general-purpose orthography-to-phonology tools.

machine-assisted transcription / segmentation automated transcription can be highly effective when trained language models exist. However, experiments on adapting available models to low-resource languages have not been promising. The CRCL wordlists supply the necessary data for an attempt to bootstrap assistive software for limited cases – e.g. recorded wordlists, which we can help locate and

provide. Similarly, the phonodynamic models we provide may give some traction to simple tasks on open audio; e.g. locating word boundaries.

minimizing resource acquisition effort we do not know how well the distribution of tokens and segments within a lexicon models typical corpus use. Nor do we know how large a subset of the lexicon is required to model the "full" (say, 10,000 words) lexicon, or how to estimate whether or not a sample in hand is sufficient. We anticipate that a combination of Monte Carlo testing, and application of Zipf's and Heaps' Laws, would address the question of devising stopping rules for minimally useful lexicon acquisition. This is a rather important question, both from the point of view of extending any of our shorter resources, and of proposing any new efforts for data acquisition (either in the field, or from untranscribed legacy field data).

evidence-based evaluation of Ethnologue / Glottolog subgrouping in comparative / historical linguistic theory, subgroups are based on objective shared phonological and lexical innovations. However, there is considerable difference between the Ethnologue and Glottolog analyses, and neither points to any clear analysis of lexical evidence. The CRCL wordlists begin to provide the data required to generate an independent subgroup analysis of languages in Asia-Pacific (based on distance measures), and to prompt the development of tools intended to specifically identify turning-point innovations. Both of these support LORELEI efforts in lexicon extension, language identification, and other language modeling applications.

linguistic data warehouse / workbench apps looking beyond the front-line performers to LORELEI tool integration, CRCL's fine-grained coverage of the Asia-Pacific region supports applications of interest to both linguists and early responders. These include the ability to project linguistic resources onto local maps, and to single out shibboleths – locally salient phonology or word forms – that help identify speakers.

7. References

Baccianella, Stefano, Andrea Esuli, Fabrizio Sebastiani. **SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining**. In Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC10) (May 2010) sentiwordnet.isti.cnr.it

Bakker, Dik, André Müller, Viveka Velupillai, Søren Wichmann, Cecil H. Brown, Pamela Brown, Dmitry Egorov, Robert Mailhammer, Anthony Grant, and Eric W. Holman. 2009. **Adding Typology to Lexicostatistics: A Combined Approach to Language Classification**. Linguistic Typology 13.167-179.

Barbançon F, SN Evans, D. Ringe and T. Warnow. 2013. **An experimental study comparing linguistic phylogenetic reconstruction methods**. Diachronica. 2013;30:143–70.

Blust, Robert. (1995). **Austronesian Comparative Dictionary (ACD)**. Honolulu: University of Hawaiʻi at Mānoa.

Blust, Robert and Stephen Trussel. 2010. **Austronesian Comparative Dictionary, web edition**. http://www.trussel2.com/ACD/.

van Dongen, S. M. (2000). **Graph clustering by flow simulation**. PhD thesis, University of Utrecht, May 2000.

Donohue, Mark, Rebecca Hetherington, James McElvenny and Virginia Dawson. 2013. **World phonotactics database.** Department of Linguistics, The Australian National University. http://phonotactics.anu.edu.au.

Dryer, Matthew S. & Haspelmath, Martin (eds.) 2013. **The World Atlas of Language Structures Online.** Leipzig: Max Planck Institute for Evolutionary Anthropology. Available online at http://wals.info

Greenhill, Simon, Robert Blust and Russell D. Gray. 2008. **The Austronesian Basic Vocabulary Database: From Bioinformatics to Lexomics**. Evolutionary Bioinformatics, 4:271-283. http://language.psy.auckland.ac.nz/austronesian

Guerini M., Gatti L. & Turchi M. **Sentiment Analysis: How to Derive Prior Polarities from SentiWordNet**. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP'13), pp 1259-1269. Seattle, Washington, USA. 2013. hlt.fbk.eu/technologies/sentiwords

Hammarström, Harald & Forkel, Robert & Haspelmath, Martin & Bank, Sebastian. 2016. **Glottolog 2.7**. Jena: Max Planck Institute for the Science of Human History. Available online at http://glottolog.org.

Jiampojamarn, Sittichai, Aditya Bhargava, Qing Dou, Kenneth Dwyer, and Grzegorz Kondrak. (2009). **DIRECTL: a Language-Independent Approach to Transliteration**. In *Proceedings of the 2009 Named Entities Workshop*, ACL-IJCNLP 2009 pp. 28-30. Suntec, Singapore.

Kamholz, David, Jonathan Pool, and Susan M. Colowick. 2014. **PanLex: Building a Resource for Panlingual Lexical Translation**. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). European Language Resources Association.

Kondrak, G. (2002) **Algorithms for Language Reconstruction**. PhD thesis, University of Toronto, Canada, 2002.

Lewis, M. Paul, Gary F. Simons, and Charles D. Fennig, Editors. 2016. **Ethnologue: Languages of the World, Eighteenth edition**. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com.

Luce, P.A. & Pisoni, D.B. **Recognizing spoken words: the neighborhood activation model**. Ear Hear. 1998 Feb;19(1):1-36.

Matisoff, James and Richard S. Cook. 2010. **On Releasing the STEDT Database.** Presentation at 43rd International Conference on Sino-Tibetan Languages and Linguistics (ICSTLL 43) at Lund University, Sweden, October 15-18, 2010.

Miller, G. A. (1995). **WordNet: a lexical database for English**. Communications of the ACM, 38(11), 39-41.

Moran, S., McCloy, D., & Wright, R. (2014). **PHOIBLE Online.** Leipzig: Max Planck Institute for Evolutionary Anthropology.

Norquest, Peter K. (2007) **A Phonological Reconstruction of Proto-Hlai**. Ph.D. dissertation. Tucson: Department of Anthropology, University of Arizona.

Ostapirat, Weera. (2000) **Proto-Kra**. *Linguistics of the Tibeto-Burman Area* 23 (1): 1-251.

Pittayaporn, Pittayawat. **The phonology of proto-Tai.** Diss. Cornell University, 2009.

Ratliff, Martha Susan. **Hmong-Mien language history**. Vol. 613. Canberra: Pacific Linguistics, 2010.

Scannell, K. P. (2007). **The Crúbadán Project: Corpus building for under-resourced languages**. In *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop* (Vol. 4, pp. 5-15).

Shorto, Harry L. (Sidwell, Paul, Doug Cooper and Christian Bauer, eds.). 2006. **A Mon–Khmer Comparative Dictionary**. Pacific Linguistics 579. Canberra: Australian National University.

Sparck-Jones, K. **A statistical interpretation of term specificity and its application in retrieval**. *Journal of Documentation*, 28(1):11-20, 1972.

Surendran, Dinoj and Partha Niyogi (2003) **Measuring the usefulness (functional load) of phonological contrasts**. University of Chicago.

Surendran, D. and Niyogi, P. (2006) **Quantifying the Functional Load of Phonemic Oppositions, Distinctive Features, and Suprasegmentals**. In O. Nedergaard Thomsen, editor, *Competing Models of Language Change: Evolution and Beyond*. John Benjamins.

Vitevitch, Michael S. Mem Cognit. 2007 **The spread of the phonological neighborhood influences spoken word recognition** Jan; 35(1): 166–175.

Warriner, A.B., Kuperman, V. and Brysbaert, M. (2013). **Norms of valence, arousal, and dominance for 13,915 English lemmas**. Behavior Research Methods, 45, pp 1191-1207

Wolff, John. **Proto-Austronesian Phonology with Glossary Vol. I.** Published by Cornell University Southeast Asia Program (SEAP) Publications, 2010.

8. APPENDIX

A MetaGloss
B MetaForm
C Phonological features
D File formats
E *Languages of Disaster* proof of concept
F Tool snapshots

Appendix A MetaGloss

MetaGloss guides the normalization of glosses. The notes below are repeated from section 8, above.

- when possible, a single WordNet 3.0 sense is provided: **house#n#1**
- when two or more useful interpretations are plausible, they are pipe-separated: **bake#v#1|toast#v#1.**
- several word classes have been added (with all items numbered #1): **d**(*emonstrative*), **j**(*conjunction*), **k**(*in term*), **m**(*odal*), **p**(*ronoun*), **q**(*interrogative*), **x** (*temporarily uncategorized*).
- when new senses are added, they are numbered #0: **armspan#n#0**.
- a polysemous sense that does not exist in English is indicated by labeling the WN 3.0 sense: **v@fist#n#1** indicates the verb sense of the noun “fist,” i.e. “make a fist.”
- kin terms are built up in regular fashion, starting with the person who is ultimately referenced: **mot.fat#k#1** is the mother of the father, or the paternal grandmother.
- senses may have *attributes* that help document what we believe is the useful reference meaning; e.g. **carry#v#1:tumpline**. This indicates that for purposes of cognate grouping the item clusters with “carry” terms, but keeps “tumpline” accessible. These head+attribute forms may be simplified in the future.
- classifiers are noted by the *:clf* attribute, e.g. **basket#n#1:clf** is a classifier for baskets, **several#a#1:clf** for several items, **kick#v#1:clf** is an instance of kicking. There may be some inconsistency in the listing of feature-oriented classifiers (e.g. long, thin items) because it is not always clear if the given form is a classifier, or just an instance of an item.
- a small amount of ad-hoc notation may be encountered, e.g.”**!**” in **!understand#v#1** negates the primary term. These affect only a few items for which proper handling is unclear.

It is important to remember that MetaGlosses do not replace the raw glosses. Rather, they provide an additional layer that is more usable as an index to phonological forms in many languages – an index that points to the forms that are most likely to be genetically related, but still respects semantic variation between lects.

Appendix B MetaForm

MetaForm guides the normalization of raw phonological transcription. Basic guidelines are simple:

- Standard IPA is always used with the exception of these characters: / *ɿ ʮ ʅ ʯ ʋ* /, which may be found in the **phonological features** table.
- Source notation that appears to indicate minor phonetic variation, and may hinder useful lect comparison, is suppressed.
- Syllable boundaries are always marked.
- Raised characters are either *diacritics* (e.g. indicating aspiration) or *secondary features* according to our analysis of the syllable.
- A *fully tokenized* form relies on three separator characters; note that tie characters are not used:
  - **¦ (x00A6 / ¦)** separates the *onset*, *nucleus*, *coda*, and *tone* sections
  - **:** separates the *core* and *post-core* sections of the onset and coda. A *pre-core* is possible, but not currently used.
  - **.** separates bound features from the pre-core and post-core, and vowels within the nucleus.
  - | separates syllables.
- A *segmented* form uses **.** to separate phonological segments.
- Some ambiguity and inconsistency are tolerated; particularly in handling of minor syllables.

Like MetaGloss, MetaForm cannot entirely replace the raw transcribed forms. Again, they help to provide an additional layer that serves as the most probable common index of features shared within and between languages.

Appendix C Phonological features

height		backness		place		manner	
high	i ⁱ y ɨ ᶤ ʉ ɯ u ɿ ʮ ʅ ʯ ʊ	*front*	i ⁱ y e ø ᴇ ɛ æ œ ɿ ʮ ʅ ʯ	*bilabial*	p b ɓ β pɸ ɸ bβ m ʙ	*nasal*	m ɱ n ɳ ȵ ɲ ŋ ɴ
near-high	ɪ ʏ ʊ	*near-front*	ɪ ʏ	*labiodental*	pf bv f ɱ v ⱱ	*stop*	p b t d ʈ ɖ ȶ ȡ c ɟ k g q ɢ ʡ ʔ
close-mid	e ø ɘ ɵ ɤ o	*central*	ɨ ᶤ ʉ ɘ ɵ ə ᵊ ɞ ɐ a ɜ	*dental*	θ ð tθ dð	*implosive*	ɓ ɗ ᶑ ʄ ɠ ʛ
mid	ə ᵊ ᴇ	*near-back*	ʊ	*alveolar*	t d ɗ ʦ ʣ s z n r ɹ ɾ l ɬ ɮ tɬ dɮ ɺ	*affricate*	pɸ bβ pf bv tθ dð ʦ ʧ ʤ ʣ ʨ ʥ tɬ dɮ tʂ dʐ cç ɟʝ kx gɣ qχ ɢʁ
open-mid	ɛ ɜ œ ɞ ʌ ɔ	*back*	ɯ u ɤ o ʌ ɔ ɑ ɒ ʊ	*retroflex*	ʈ ɖ ᶑ ʂ ʐ tʂ dʐ ɳ ɽ ɻ ɭ	*fricative*	ɸ β f v θ ð s z ʂ ʐ ʃ ʒ ɕ ʑ ç ʝ x ɣ χ ʁ ħ ʕ h ɦ ɬ ɮ
near-low	æ ɐ			*palatoalveolar*	ʃ ʒ ʧ ʤ	*approximant*	w ɹ ɻ j ɰ ɥ l ɭ ȴ ʎ ɫ
low	a ɑ ɒ			*alveolopalatal*	ȶ ȡ ɕ ʑ ʨ ʥ ȵ ȴ	*tap-flap*	ⱱ ɾ ɽ ɺ
				palatal	c ɟ ʄ ç cç ɟʝ ʝ ɲ j ʎ	*trill*	ʙ r ʀ ʜ ʢ
				labiopalatal	ɥ		
				velar	k g ɠ kx gɣ x ɣ ŋ ɰ ɫ		
				labiovelar	w		
				uvular	q ɢ χ ʁ qχ ɢʁ ɴ ʀ ʛ		
				pharyngeal	ʡ ħ ʕ ʜ ʢ		
				glottal	ʔ h ɦ		

class		role		position		length		pre-articulation	
consonant		*onset*		~~*pre*~~		*epenthetic*	ᵊ ⁱ ᶤ	*prenasalized*	ᵐ ⁿ ᶯ ᶮ ᵑ ᶰ
vowel		*nucleus*		*core*		*short*	x̆	*devoiced*	x̥ x̊
syllabic	x̩	*coda*		*post*		*long*	ː	*preglottalized*	ˀ
minor				*1.1 1.2 2.2 1.3 2.3 3.3 1.4 2.4 3.4 4.4*				*preaspirated*	ʰ
								prelabialized	ʷ
								prestopped	ᵇ ᵈ ᶡ ᶢ

realization		phonation		post-articulation	
rounded	ɥ ʮ ʋ y ø œ ʉ ɵ ɞ ɑ ɔ o u ʏ	*nasal*	x̃	*nasalized*	ᵐ ⁿ ᶯ ᶮ ᵑ ᶰ
voiced	m ɱ n ɳ ɳ̊ ɲ ɴ b d ɖ ɟ ɡ ɢ ʡ bβ bv dð ʤ ʥ ʥ dʒ dʑ ɟʝ ɡɣ β v ð z ʐ ʒ ʝ ɣ ʁ ʕ w r ɹ ɾ ɻ j ɰ ɥ l ɮ ʎ ɫ ɭ̊ ɭ ɺ ʙ ʀ ʜ ʢ ɢʁ ʑ ȡ ɦ	*aspirated*	ʰ	*glottalized*	ˀ
retroflexed	ɻ ɻ	*devoiced*	x̥ x̊	*palatalized*	ʲ
lateralized	l ɬ ɮ tɬ dɮ ɺ ɭ ʎ ɫ ɭ̊	*breathy*	x̤	*labialized*	ʷ
fricated	ɿ ʯ ʅ ʮ ʋ	*creaky*	x̰	*labiopalatalized*	ᶣ
nonvocalized		*dental*	x̪	*stopped*	ᵇ ᵈ ᶡ ᶢ
prevocalized		*raised*	x̝	*velarized*	x̴ ˠ
vocalized		*lowered*	x̞	*pharyngealized*	xˤ
		rhotic	x˞		

Appendix D File formats

Files discussed below exemplify the full distribution. When appropriate, the /Ethnologue path and files are paralleled by a /Glottolog set (and may be expanded to other analyses). Below, the **#File:** line (giving the path) is not part of the file. Commented lines in **bold** text are column labels.

```
#File: crcl/paths.txt
#File: crcl/geo/info.geo
#File: crcl/geo/CN/info.geo
#File: crcl/geo/CN/Yunnan.geo
#File: crcl/metagloss/metagloss.txt
#File: crcl/metagloss/new.txt
#File: crcl/metagloss/kin.txt
#File: crcl/metagloss/n.txt
#File: crcl/cognates/setByRow.tsv
#File: crcl/cognates/setByCol.tsv
#File: crcl/AA/alk/huffman1971vocabulary.c12/2_segment.tsv
#File: crcl/AA/alk/huffman1971vocabulary.c12/3_segment_nuc.tsv
#File: crcl/AA/alk/huffman1971vocabulary.c12/cover.tsv
#File: crcl/AA/alk/huffman1971vocabulary.c12/contrast.tsv
#File: crcl/AA/alk/huffman1971vocabulary.c12/density.tsv
#File: crcl/AA/alk/huffman1971vocabulary.c12/load.tsv
#File: crcl/AA/alk/info/Ethnologue/geo_distance.tsv
#File: crcl/AA/alk/info/Ethnologue/ety_distance.tsv
#File: crcl/AA/alk/info/Ethnologue/ety_lexicon.tsv
#File: crcl/AA/alk/info/Ethnologue/geo_lexicon.tsv
#File: crcl/AN/mak/text/Scannell/mak-Latn/info.txt
#File: crcl/AN/mak/text/Scannell/mak-Latn/urls.txt
#File: crcl/AN/mak/text/Scannell/mak-Latn/chartrigrams.txt
#File: crcl/AN/mak/text/Scannell/mak-Latn/wordbigrams.txt
#File: crcl/AN/mak/text/Scannell/mak-Latn/words.txt

#File: crcl/cognates/cognates.xml
#File: crcl/cognates/etygloss/able#a#1/AA:S1179.xml
```

```
#File: crcl/paths.txt
   #path
   crcl
   crcl/paths.txt
   crcl/hadr
   crcl/metagloss
```

Paths to all files.

```
#File: crcl/geo/info.geo
   #bibref column            ISO    country        ADM-1                          lat,long
   tryon1995comparative 80   rap    Chile                                         -27.1248,-
   109.3571
   hudak2008comparative 1    tha    Thailand       Changwat Lop Buri
        14.7368,100.5249
   hudak2008comparative 5    tts    Thailand       Changwat Maha Sarakham
        16.1155,102.9990
   hudak2008comparative 8    nod    Thailand       Changwat Lampang
        18.3471,99.7262
```

The top-level list only provides Ethnologue data because it has slightly better ISO 639-3 coverage. Latitude and longitude are typically 4-digit reals, and reflect the location of the populated place nearest to the *lat,long* figure we license from SIL, and which cannot be released.

```
#File: crcl/geo/CN/info.geo
   #bibref column      ISO     country ADM-1          lat,long
   huang1992tbl 10     pmi     China   Sichuan Sheng  27.9014,101.5165
   huang1992tbl 11     jya     China   Sichuan Sheng  31.7580,102.2552
   huang1992tbl 12     ero     China   Sichuan Sheng  30.8187,101.8259
   huang1992tbl 13     qvy     China   Sichuan Sheng  30.3193,100.8392
```

These have the same format as the top-level **crcl/geo/info.geo** file. The country code is the two-letter ISO 3166-1 alpha-2 abbreviation. The summary **info.geo** file is provided because ADM-1 codes cannot always be identified for a given *lat,long* value (e.g. if it happens to fall in open water). We expect to resolve these over time.

```
#File: crcl/geo/CN/Yunnan.geo
   #bibref column      ISO     country ADM-1          lat,long
   huang1992tbl 20     duu     China   Yunnan         27.9801,98.4442
   huang1992tbl 28     acn     China   Yunnan         24.6798,98.7253
   huang1992tbl 29     acn     China   Yunnan         24.6798,98.7253
   huang1992tbl 30     atb     China   Yunnan         24.4029,98.3244
```

These have the same format as the top-level **crcl/geo/info.geo** file, and describe the current ADM-1.

```
#File: crcl/metagloss/metagloss.txt
   #metagloss  count
   a           592
   d           11
   j           7
   k           159
```

The **metagloss.txt** file summarizes the POS-specific files; however, they split all ***a|b*** forms into the individual words (which may have different POS).

```
#File: crcl/metagloss/new.txt
   #metagloss          count   explanation
   a_little#n#0        38
   among#r#0           30
   armspan#n#0         121
   armspan#n#0:around  41
```

The top-level list only provides Ethnologue data because it has slightly better ISO 639-3 coverage. Latitude and longitude are typically 4-digit reals, and reflect the location of the populated place nearest to the lat,long figure we license from SIL, and which cannot be released.

```
#File: crcl/metagloss/kin.txt
   #All kin term components in use
   .BY. address term:  a.BY.b
   Post-modifiers
   :addr       general address term
```

This file documents the construction of kin terms in MetaGloss.

```
#File: crcl/metagloss/n.txt
   #POS              count
   1#n#1             37
   Adam's_apple#n#2  35
   Allium#n#1        5
   April#n#1:lunar   36
```

This particular file lists all noun forms that appear in MetaGloss. Other x.txt POS files are similar: **a**:adjective, **d**:demonstrative, **j**:conjunction, **k**:kin, **m**:modal, **n**:noun, **p**:pronoun, **q**:interrogative, **r**:adverb, **v**:verb, **x**:unassigned

```
#File: crcl/cognates/setByRow.tsv
   #count  EtySet          cogset            arnaud1997lexique.c1   arnaud1997lexique.c2
           arnaud1997lexique.c3 ...
   7       Allium#n#1      HM:R599
   9       Allium#n#1      HM:R835
   15      Hmong#n#1       HM:R73
   17      I#p#1           AA:2
```

Each **EtySet** is the rough gloss of a historical form, while each **cogset** includes related terms from modern languages, given in the appropriate cell (most cells are empty). We expect there to be at least one cogset per family. Cogsets are named either by a reference to the literature, or by an arbitrary number associated with the family. Over time, both cogsets and etysets will cluster into larger groupings of genetically related forms.

```
#File: crcl/cognates/setByCol.tsv
   #count  source                 ISO      Allium#n#1|HM:R599     Allium#n#1|HM:R835
           Hmong#n#1|HM:R73       ...
   220     arnaud1997lexique 10   npy
   391     arnaud1997lexique 11   sda
   369     arnaud1997lexique 12   mqj
   262     arnaud1997lexique 14   rog
```

The *setByCol* view labels each column with an **EtySet|cogset** pair. The **count** gives the number of items from a particular source have been assigned to cogsets. These items appear in the table cells (most are empty). Over time, cells will contain more forms as cognate sets are first developed following current semantics, then joined to account for semantic shift and borrowing.

```
#File: crcl/AA/alk/huffman1971vocabulary.c12/2_segment.tsv
   #huffman1971vocabulary 12  AA     alk     Guibian Zhuang
   <    k       90
   h    >       88
   ŋ    >       87
   <    t       78
```

Segment bigrams and counts. Pre- and post-word boundaries are shown with < and >. The first line gives the table contents: bibref and column, family, ISO 639-3 code, and ISO language name.

```
#File: crcl/AA/alk/huffman1971vocabulary.c12/3_segment_nuc.tsv
   #huffman1971vocabulary 12   AA      alk     Guibian Zhuang
   <    k       a       42
   <    p       a       31
   <    t       a       28
   <    pʰ      a       26
```

Segment bigrams, as above, except that the complete nucleus (diphthongs and longer) is treated as a single segment. Other 2_..., 3_... files are similar, with content as per file name.

```
#File: crcl/AA/alk/huffman1971vocabulary.c12/cover.tsv
   #huffman1971vocabulary 12   AA      alk     Guibian Zhuang
   #64 letters, 31 words
   #  a aː b c cʰ d e eː f h i iː j k kʰ kʷ l m m̩ n n̩ o oː p pʰ r rʷ s t tʰ u uː w ŋ ŋ̊ ŋ̩ ɑ ɔ ɔː ə
   əː ɛ ɛː ɨ ɨː ɬ ɲ ɲ̩ ʔ ʔj ʔl ʔr ʔw ᵐb ᵐp ᵐpʰ ᵑk ᵑkʰ ᶮc ᶮcʰ ⁿt ⁿtʰ
   pruŋ tɨp kasɔk      grave#n#2
   tʰaluːp tʰanəːj     clothing#n#1
```

Minimum cover set. Line 1 describes the source and language. Line 2 gives the number of distinct phonological segments, and the size of the minimum cover set. The remainder of the file consists of (tab-separated) words and their glosses.

```
#File: crcl/AA/alk/huffman1971vocabulary.c12/contrast.tsv
   #huffman1971vocabulary 12   AA      alk     Guibian Zhuang
   a      aː     kat     kaːt     k#t     from#r#0/only#a#1     burn#i#3
   a      aː     paj     paːj     p#j     three#n#1             rice#n#1:cooked
   a      aː     pʰat    pʰaːt    pʰ#t    grass#n#1             chew#v#1
   a      aː     tap     taːp     t#p     stab#v#2              slap#v#1
```

Minimum contrast set. The columns show the two contrasting segments, the words each appears in, and a joint form with # in the common slot. The final columns have the metaglosses of the two contrasting words.

```
#File: crcl/AA/alk/huffman1971vocabulary.c12/density.tsv
  #huffman1971vocabulary 12
  #Clustering coefficient (2Nv/Kv(Kv-1))   Links (Nv)   Degree (Kv)   word     neighbors
  0.7778                                   28           9             caː
                                           cɔː|jaː|kaː|maː|naː|raː|taː|tʰaː|ᵐpʰaː
  0.3611                                   13           9             maj
                                           mat|maŋ|maʔ|moːj|paj|saj|ʔaj|ᵑkaj|ᶮcaj
  0.3333                                   12           9             paj
                                           par|pat|paŋ|paɲ|paːj|saj|ʔaj|ᵑkaj|ᶮcaj
```

Each list of neighbors differs from the target word by a single phonological segment. **Kv** is the number of these neighbors. **Nv** is the number of neighbors that are one segment away from each other. The **clustering coefficient** is in the range 0 .. 1, and gives a sense of how tightly bound the neighborhood is.

```
#File: crcl/AA/alk/huffman1971vocabulary.c12/load.tsv
  #huffman1971vocabulary 12   AA      alk     Guibian Zhuang
  #segment      contrst total   load
  a     35      378     0.0925
  aː    29      101     0.2871
  b     11      6       1.8333
```

Segment bigrams, as above, except that the complete nucleus (diphthongs and longer) is treated as a single

segment. Other 2_..., 3_... files are similar, with content as per file name.

```
#File: crcl/AA/alk/info/Ethnologue/geo_distance.tsv
  #ISO  analysis       0-100      101-200    201-300     301-400        401-500
  alk   Ethnologue
        llo:10|oyb:15|irr:17|ngt:26|spu:32|lbo:41|skk:49|tto:49|nev:51|kuf:54|tth:64|tgr:74|kgd:8
  0|oog:80|jeg:83|kgc:85|sqq:97
        stg:103|pac:107|hld:112|brb:113|phg:114|ktv:116|brv:121|tdf:121|jeh:137|krv:151|hal:151|t
  kz:154|bru:158|rmx:163|ren:169|sed:174|xkk:192|tdr:195|kta:196|cua:199
        kxy:209|xhv:216|moo:217|krr:221|tpu:228|sss:237|hre:240|jra:242|yoy:245|nuo:252|nyl:255|s
  cb:256|bdq:261|pcb:267|skb:287|kdt:295|pkt:296
        rka:305|uan:312|aem:312|nyw:314|cmo:321|pht:327|vie:340|rad:343|thm:343|bgl:354|kxm:362|h
  ro:365|tmp:369|bfk:384|lso:385|tts:390|tpo:393
        mng:406|khm:407|mnn:407|hnu:412|cja:417|sti:418|tnu:419|huq:422|stt:427|tyj:437|tou:439|r
  og:443|cma:455|kpm:463|cuq:469|lic:474|cje:479|syo:483|jio:483|tyh:488|tmm:489|thc:492|lao:495
```

Each row is labeled with the current ISO-639-3 code and the source (Ethnologue or Glottolog) of the language position points. The remainder of the row has five tab-separated groups of *ISO:distance* pairs, each | separated. Distances are in kilometers, using single-point language locations; these are progressively less meaningful as the size of the speaker community increases. A future release will attempt to take national or regional languages into account, regardless of their point distance. ISO codes are only given for languages we have data for.

```
#File: crcl/AA/alk/info/Ethnologue/ety_distance.tsv
  alk   Ethnologue     alk|tpu
        alk|brb|bru|brv|cbn|cog|irr|jeh|kdt|kgc|kgd|khm|kjg|ktv|kuf|lbo|lcp|mlf|mnw|ngt|oog|pac|p
  cb|sss|sti|tdf|tpu|tth|tto
        alk|brb|bru|brv|cbn|cog|irr|jeh|kdt|kgc|kgd|khm|kjg|ktv|kuf|lbo|lcp|mlf|mnw|ngt|oog|pac|p
  cb|sss|sti|tdf|tpu|tth|tto
```

Each line has three tab-separated groups of ISO codes that share the same parent, grandparent, and great-grandparent; note that some groups may be identical. Ethnologue data is (for the moment) suspect due to problems in properly identifying some parent levels. Only groups with <=50 ISO codes are reported, because a single early-branching survivor (common in Austronesian) may include the entire family as first cousins. ISO codes are only given for languages we have data for. The Glottolog analysis tends to have more branches / smaller groups.

Each ASJP line lists ISO codes and the NDLD (*normalized Levenshtein distance divided*) from the current ISO code [Bakker et al 2009]. A maximum of 50 codes are provided. In some cases a distance from the current ISO code (to itself) may be reported; this occurs when the ASJP dataset had multiple lect samples.

```
#File: crcl/AA/alk/info/Ethnologue/ety_lexicon.tsv
  #sources  gloss      [alk] huffman1971vocabulary 12       [alk] theraphan2001languages_2 4
            [tpu] huffman1971vocabulary 16 ...
  2         Idau#k#1   c.a.w
            k.ɨ.m.n.a.n k.a.n
  3         Ifat#k#1   t.aː                                 t.aː
            pʰ.ɨ.ʔ n̥ə.j
  3         Imot#k#1   j.a.ʔ                                j.a.ʔ
            m.a.e.ʔ n̥ə.j
  2         Ison#k#1   c.a.w                                p.a.s.aː.w
```

A lexicon of sister languages according to a specific subgroup analysis Trees for Ethnologue and Glottolog are similar. Each row is labeled with the number of lects that have forms for the gloss in the second column. All entries in each sister-language lexicon is included; however, some rows may just have a single form entry.

```
#File: crcl/AA/alk/info/Ethnologue/geo_lexicon.tsv
  #sources   gloss                [AA:alk:0] huffman1971vocabulary 12  [AA:alk:0] theraphan2001languages_2 4
             [AA:irr:17] huffman1979vocabulary 1 ...
  2          !understand#v#1                                           c.ɔː.m

  3          Careya_arborea#n#0   k.a.d.oː.n
  3          Caryota#n#1                                               t.a.j.uː.ŋ
  4          Hypericacaea#n#1                                          h.a.ŋ.i.a.ŋ
             h.a.ŋ.i.ə.ŋ
```

A lexicon of lects whose point locations are within 100km of each other. Trees for Ethnologue and Glottolog are similar. **NB:** this is a wide table; the data values shown here are for illustrative purposes only.

```
#File: crcl/AN/mak/text/Scannell/mak-Latn/info.txt
  ISO 630-3     mak
  BCP-47        mak-Latn
  glottocode    maka1311
  name          Makasar
  country       Indonesia (Sulawesi)
```

The Scannell info file summarizes the per-BCP-47 code information he provides.

```
#File: crcl/AN/mak/text/Scannell/mak-Latn/urls.txt
  http://incubator.wikimedia.org/wiki/Wp/mak/Gowa
  http://incubator.wikimedia.org/wiki/Wp/mak/Main_Page
  http://incubator.wikimedia.org/wiki/Wp/mak/Persigowa_Gowa
  http://incubator.wikimedia.org/wiki/Wp/mak/PSM_Mangkasara%27
  http://www.bible.is/toc?version=MAKLAI&language=Makassar
```

Scannell source file; shows links to his data sources.

```
#File: crcl/AN/mak/text/Scannell/mak-Latn/chartrigrams.txt
  ang 25834
  ng> 15144
  ri> 15101
  na> 13937
  <an 12413
```

Scannell source file; contains character triples and counts. < and > indicate word boundaries.

```
#File: crcl/AN/mak/text/Scannell/mak-Latn/wordbigrams.txt
  . \n 10623
  mae ri 2290
  , " 1820
  . " 1021
  " \n 1015
```

Scannell source file; these are space-separated token bigrams and counts.

```
#File: crcl/AN/mak/text/Scannell/mak-Latn/words.txt
  ri 11702
  anjo 5001
  siagang 3687
  ke'nanga 3110
  Allata'ala 3031
```

Scannell source file; these are space-separated tokens and counts.

```
#File: crcl/cognates/cognates.xml
   <document version="1.0">
     <cognate id="huffman1971vocabulary:C:c13.r625.gs2041.i8527" iso639-3="lbo" lang="Laven">
       <etygloss>roast#v#1</etygloss>
       <cogset>AA:S2041</cogset>
       <form>buh</form>
     </cognate>
   ...
   <cognate id="huffman1971vocabulary:C:c9.r625.gs2041.i8526" iso639-3="kdt" lang="Kuy">
   <etygloss>roast#v#1</etygloss> <cogset>AA:S2041</cogset> <form>buh</form></cognate>
   <cognate id="huffman1979vocabulary:C:c10.p19-29.r1471.i11948" iso639-3="sss" lang="Sô">
   <etygloss>roast#v#1</etygloss> <cogset>AA:S2041</cogset> <form>buh</form></cognate>
   <cognate id="huffman1979vocabulary:C:c11.p21-29.r1474.i14776" iso639-3="tto" lang="Lower
   Ta'oih"> <etygloss>roast#v#1</etygloss> <cogset>AA:S2041</cogset> <form>boh</form></cognate>
```

The complete set of cognate entries. The **cognate** tag encapsulates each entry, with attributes *id* (consistent across all data), an *iso639-3* code, and the formal ISO *lang* language name. The **etygloss** gives a rough historical semantic label; each **cogset** numbers a cognate set. The form (like the attributes) are included for convenience, and can be recaptured from the main dataset. **NB:** The entry has been indented for display.

```
#File: crcl/cognates/etygloss/able#a#1/AA:S1179.xml
   <document version="1.0">
     <cognate id="huffman1971vocabulary:C:c1.r39.gs1179.i641" iso639-3="khm" lang="Central
   Khmer">
       <etygloss>able#a#1</etygloss>
       <cogset>AA:S1179</cogset>
       <form>ba:n</form>
     </cognate>
```

Identical to the same item in the complete cognate set, above.

Appendix E: Languages of Disaster

CRCL proposes to build a resource that locates, enriches, and ties language and GIS data to humanitarian assistance / disaster relief (HA/DR) event histories. It will support applications for responding to, and predicting or pre-provisioning, disaster events. We attempt to balance today's desire for an interactive sandbox with tomorrow's probable request for machine access to data for re-use and/or re-implementation. This document describes the project's goals, content, and development issues. An initial proof of concept can be found at http://sealang2.net/project/lorelei/over.

Introduction

The DARPA LORELEI project is based on the observation that language information is integral to effectively detecting, directing, and delivering HA/DR assistance. Most of the current research effort frames the issue from the point of view of *response* that involves given *target* languages: how can we most effectively analyze communications in a particular language in a disaster situation?

We extend this by considering the issue from the point of view of both response to and anticipation of events. Given an impending disaster, what geographic areas and speaker communities will be affected? Given a history of disaster characteristics (frequency, duration, extent, impact), as well as understanding of language distributions and relations, can we predict not only what areas and communities a particular kind of disaster will effect, but also what languages might be most usefully pre-provisioned? This information is helpful to both users and providers of LORELEI capability.

We also consider the problem from the distinct viewpoints of LORELEI 1.* *performers*, and the analysts who are our ultimate downstream *consumers*. For example the performer wants to know the likely source of loan words into a target language; the analyst want to know what language(s) a random person in an arbitrary city is likely to speak. The performer wants an aggregate model that helps in machine-based language identification, while the analyst needs to know likely forms for "hungry" within a 10-mile radius.

A secondary goal of the project is to make the somewhat inchoate mass of language-relevant information more discoverable and comprehensible. We want to be able to instantly answer such questions as: is MT technology available for a given language? What is the most similar language that has either MT, or substantial data resources? Are text samples available? If not, what wider language of communication is likely to have influenced a given language's writing system? What related and unrelated languages inhabit the same general geographic area, and what are their relative speaker numbers?

Considerable work has been done on each of this problem's three major aspects: linguistics, geodata, and HA/DR data. Unfortunately, we cannot produce a useful tool simply by mashing datasets together; there are non-trivial problems to solve in both harmonizing and extracting actionable information from the data. By the same token, even given harmonized, mashable data, it is not instantly clear what the most effective ways to articulate queries and display results should be. We built the proof-of-concept website to explore this question.

Design principles

Our first premise is that any one or more of three basic parameters – languages, HA/DR events, and geographic areas – should be able to serve as a search key for any of the others. This is achieved by indexing each data set in terms of one or more ADM-1 top-level administrative areas, typically provinces or states. Thus, the ADM-1 is the common key to all data.[2] Implicitly, features of any one set link to the others via the ADM-1; e.g. a date range implicitly links to languages effected by HA/DR events that fall in that range, and effect those speaker areas.

Second, we want to be able to aggregate results whenever possible. A data-driven choice of a high-value *investment language* – that is, a language that should arguably be pre-provisioned – depends on understanding not only its similarity to related languages and the availability of existing resources, but also the expected impact of future disasters on speaker communities which might benefit. We cannot assume that well-provisioned national languages (such as Thai or Vietnamese) will fill this role, incidentally – their very success (and the integration of foreign influences this usually implies) often makes these languages poor examples of the family or branch as a whole.

Third, we try to anticipate and enable any logical needs for follow-through / drill down / loop back. For example, a query into events that have affected a region will return mentions of countries and languages. The natural drill-down is to click on one of these countries or languages in order to see what events have impacted it. Then, we're likely to want to loop back – click on an event to re-use it as a new starting query – because it delimits a region or set of languages.

Fourth, we're interested in what might be called *analytical imagery*. The demo site shows some simple examples of how weighting can be used to render maps that may help clarify unseen relations, such as the contrast between the number of events, and their impact in terms of population and speaker community numbers.

Finally, we want to *expose data*, and not just analyzed results, in support of decision making. Our goal is not to replace the analyst, but rather to provide all available information, allowing alternative views of single data sets, and comparison of alternative data sets. We also want to allow drill-down into any of the language / event / geographic axes, e.g. a visual interface may be useful for discovery, but a lexical dataset, list of languages by city, or contemporaneous news reports might ultimately be most useful to the analyst. Thus, we anticipate providing machine access to data.

Data sources are discussed in more detail below, but briefly:

- disaster data is taken from the EM-DAT and GLIDE datasets,
- GIS data is from the GADM shapefile sets, which attempt to cover all five ADM levels worldwide, and GeoNames.org, which has the best vernacular and informal name information,
- linguistic data is from a variety of sources: subgrouping from Ethnologue, Glottolog, and ASJP, MT availability from our own survey of Google, Bing, and Yandex resources, base-level resource availability from GlottoDoc, corpus availability from An Crubadan,
- secondary data is inferred whenever possible.

---

[2] This also turns out to be an effective granularity from the linguistic perspective – ADM-1 boundaries are not necessarily arbitrary political boundaries; rather they often delimit geographic, ethnic, and linguistic areas.

The current implementation takes a few shortcuts. For example, the GLIDE data is only roughly integrated (it will ultimately be tied to EM-DAT, which has much better geographic extent data). And we use each language's nominal center point to identify a single ADM-1 entity (in fact, it might be spoken in several). We can sometimes mitigate these; for example, GLIDE data can be roughly aligned by incident date, and a country's national language(s) can be assigned to every ADM-1.

Functionality and use cases

To varying degrees, CRCL's */over* website provides the following types of information and functionality:

- resource availability for all 7,100 living languages per the ISO 639-3 standard,
- resource availability within LORELEI,
- impact of disasters on speaker communities,
- the likely national and regional second/third languages for each speaker community,
- the 'nearest' (per Ethnologue/Glottologue) relative that has tools or large data resources.
- the condition and reliability of state-of-the-art disaster and speaker data.
- various maps that show weighted event distributions,
- a summary of event types and languages affected,
- analysis of likely high-value investment language candidates.

Typical use cases for 1.* LORELEI developers include identification of:

- suitable incident languages, which have an appropriate mix of population and existing resources.
- high-value investment languages – those that are directly or indirectly the target, fallback, or pivot for – high-risk regions,
- languages, regions, and dates of known past events, which may be used to help model and recognize on-line "disaster chatter."

Finally from the analyst's perspective, we can explore:

- impact of past events of the same type in the same area,
- speaker communities likely to be affected, and their populations,
- languages likely to be used/understood in each city in the area,
- language resources available,
- most likely broad language(s) of communication,
- external reports linked to the EM-DAT or GLIDE identifiers (not yet implemented),
- a set of HA/DR query terms for each language (e.g. CRCL's HA/DR parallel lexicon sets),
- ideally, a model of historical "disaster chatter" esp. in languages that are not currently modeled or discoverable (I like the HA/DR lexicon, but I'm not convinced that it can properly seed for or identify all relevant online data).

Disaster Resources

The primary disaster resources are **EM-DAT** and **GLIDE**. Both provide numbers that identify event type and date. A separate number is issued for each country; i.e. a single event may have multiple numbers.

GLIDE The *Global Identifier Number* system was developed by the *Asian Disaster Reduction Center* (ADRC). It includes 6,259 event references (http://glidenumber.net). GLIDE supplies somewhat longer text descriptions of the events, as well as a single latitude / longitude point (derivation is unclear).

EM-DAT The *Emergency Events Database*, produced by the *Centre for Research on the Epidemiology of Disasters* (CRED). "EM-DAT contains **essential core data** on the occurrence and effects of over **22,000 mass disasters** in the world from **1900 to the present day.** The database is compiled from various sources, including UN agencies, non-governmental organizations, insurance companies, research institutes and press agencies." (http://emdat.be)

EM-DAT supplies a text description of each numbered event's area. In theory this is an administrative area as specified by **GAUL** (discussed below), but in practice locations are given as a mix of formal and informal names. In some cases, EM-DAT also provides estimates of the financial impact, number of deaths, and number of people affected by each event.

Both GLIDE and EM-DAT numbers are sometimes cited in other databases. However, regular citation (a la ISBN numbers) is not common.

Shortcomings GLIDE and EM-DAT are not cross-linked. Because they do not always record events as occurring in the same time or place, they will require a combination of machine and hand alignment. While GLIDE's lat/long points are helpful for obtaining a quick visual overview of events in a region, they given no indication of the actual extent of any event. EM-DAT does a much better job of listing affected areas; however, the public dataset does not normalize these names to GAUL ADM-1 names. Again, we can do quite a bit of heavy lifting by machine, but hand alignment will also be required.

As noted, the EM-DAT impact estimates are incomplete. We will provide parameters for estimating the blanks by using known relations between cost/death/affected figures, and between known impacts and event types.

Geo Data Resources

Primary resources are listed here. We rely on GADM, with additional support from GeoNames.

GADM The *Global Administrative Areas* project provides shapefiles for all five ADM levels for all countries. It currently has data for 294,430 administrative areas. This is the best open shapefile source, and has reliable ADM identification. (http://gadm.org)

GeoNames This is the most extensive set of place names and equivalents available. "The GeoNames geographical database ... contains over 10 million geographical names and consists of over 9 million unique features whereof 2.8 million populated places and 5.5 million alternate names. All features are categorized into one out of nine feature classes and further subcategorized into one out of 645 feature codes." (http://geonames.org/about.html). Lat/long points are provided for each item.

GNS The *Geographic Names System* (US National Geospatial-Intelligence Agency) set is the US standard. It only includes point information for ADM-1 entities. (http://geonames.nga.mil/gns/html/)

GAUL The *Global Administrative Unit Layers* dataset is prepared by the United Nations / FAO. It includes shapefiles for ADM-1 and ADM-2 entities. It is not publicly available, however, there is a released crosswalk to GNS. See

http://www.fao.org/geonetwork/srv/en/metadata.show?id=12691 and http://blog.gdeltproject.org/global-second-order-administrative-divisions-now-available-from-gaul/ .

Shortcomings The resources above reflect the distinctive primary concerns of their developers, and it is probably better to think in terms of each set's strengths rather than its weaknesses. GeoNames is extremely helpful for indentifying non-standard and vernacular names, but names may be missing, or under-specified (in term of ADM category). GADM has excellent coverage of formal names, and has both points and polygons, but is not sufficient for identifying place names found in the wile.

As noted above, because place naming in EM-DAT is somewhat irregular, normalizing its combination of (usually) ADM-1 and ADM-2 names to GADM will require a combination of machine and hand work.

Language Data Resources

Subgrouping To determine language similarity globally we rely on **Ethnologue, Glottolog**, and **ASJP** (the *Automated Similarity Judgment Project*). All use ISO 639-3 codes for language indexing; however, Glottolog rejects some of these and maintains a parallel set (*glottocode*) of finer-grained lect-by-lect identifiers. Ethnologue and Glottolog provide roughly the same family and subgroup analyses; however, Glottolog tends to split (and Ethnologue tends to lump) lower-level sub-branches. ASJP does not provide a branch analysis per se; rather, one can build a table of distance measures for all languages.

This is an area in which data and methodology from CRCL and other LORELEI performers should be able to make a significant improvement. The Ethnologue and Glottolog analyses are based on (sometimes idiosyncratic) interpretations of what constitutes a significant phonological innovation; this does not always speak to similarity from the point of view of machine translation or language identification. ASJP uses tiny (40-item) sets; these may distinguish the major family and branch splits, but are less effective at finer levels.

Machine Translation We list the open access tools provided by *Google*, *Bing*, and *Yandex*, including development languages, as a proxy for the availability of "advanced" language technology resources. These languages probably have other necessary resources (text and bitext corpora, dictionaries) available.

Text Corpora As noted above, available MT resources usually predict corpus availability for the major languages. For the other 98%, Scannell's *An Crubadan* is believed to be the broadest corpus set known.

Demographic Data We license the *Ethnologue 18* dataset. This provides speaker number approximations, and details regarding each language's official status (which is helpful for inferring secondary languages of communication). Speaker area data is based on Ethnologue; see e.g. http://langscape.umd.edu/map.php. We will not distribute any shapefile data.

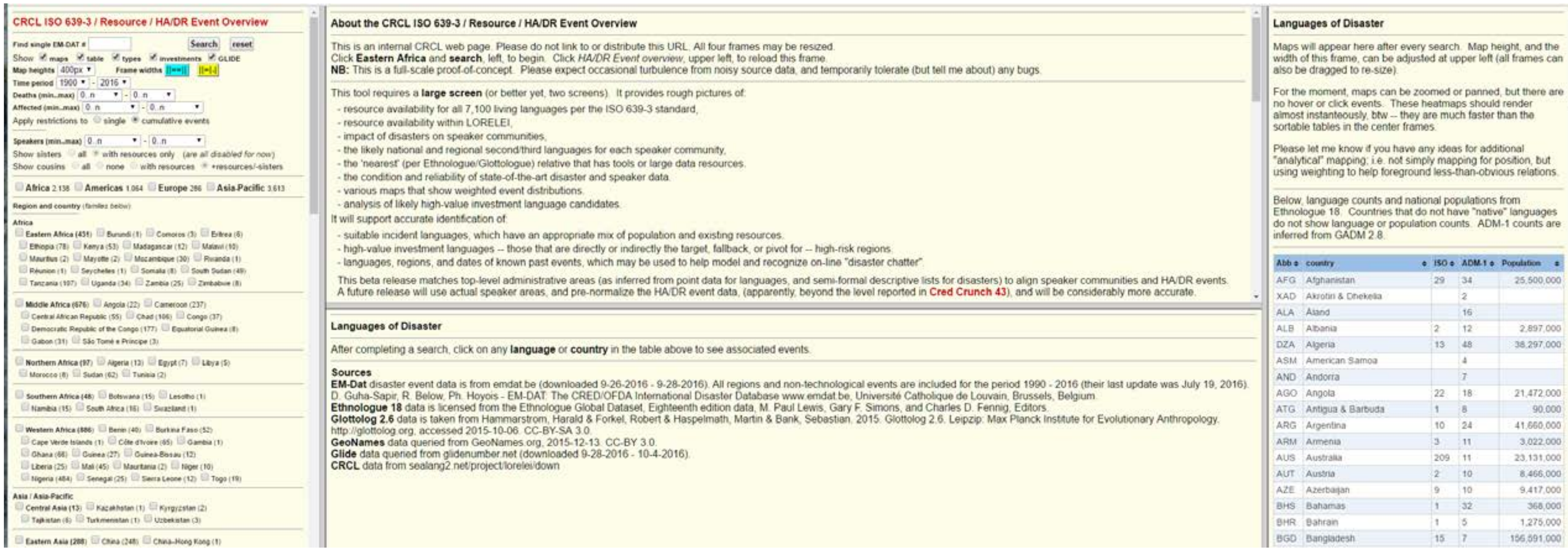
Proof-of-concept

An initial proof of concept can be found at http://sealang2.net/project/lorelei/over. It demonstrates most of this proposal's ideas, but still requires work in various areas:

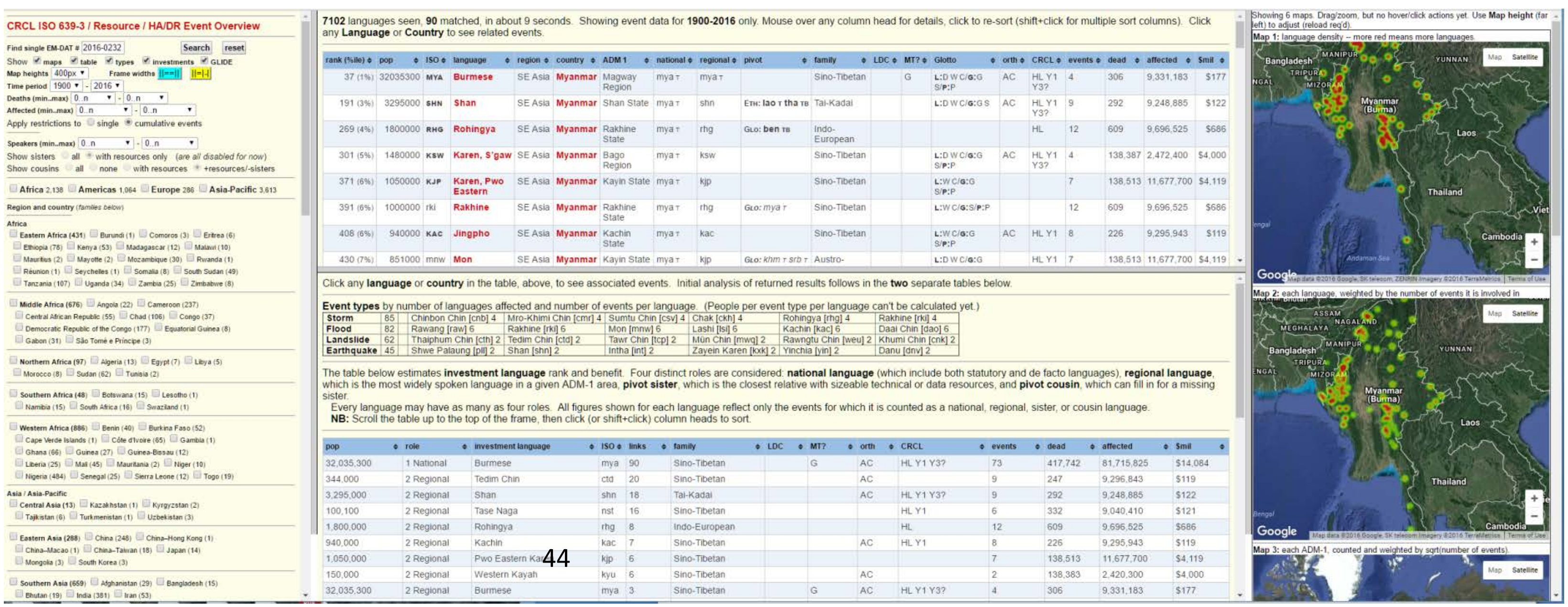
- documenting website functionality,
- aligning the GLIDE and EM-DAT event numbers,
- revising the EM-DAT location data to reflect precise ADM entities,
- obtaining city and ADM-1-level data on language distribution,
- parameterizing measures that estimate missing death, damage, and affected population figures,
- improving the current quick-and-dirty similarity measures used to identify pivot languages,
- adding mouse functionality to the map displays,
- providing additional functionality for summarizing historical events by language(s) and vice versa,
- linking CRCL's "small lexicons," and very large set of HA/DR parallel lexicons, to the interface,
- identifying and providing click-through access to other external data sources that are accessible via EM-DAT and/or GLIDE numbers.

Annotated screen captures follow, below.

Complete browser: menu, center top, center bottom, maps



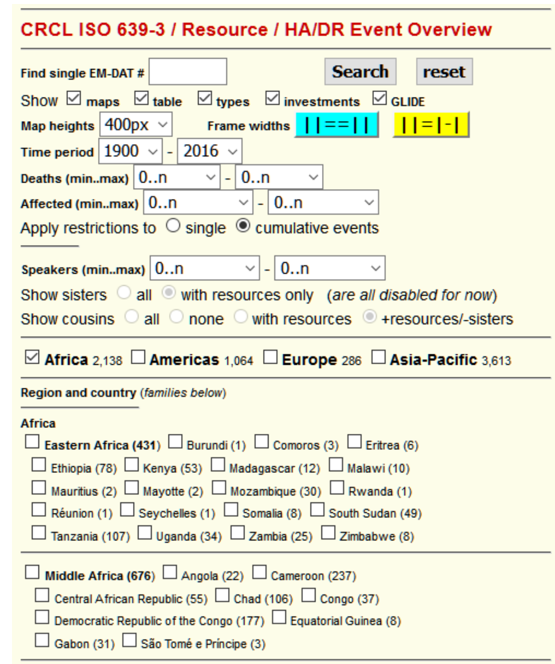
Above, the initial site view. The capture below was taken after a single query, selecting only "Myanmar" in the menu on the left. Note that the center frame has separate top and bottom portions – the top contains a sortable table, and the bottom has a fixed table, followed by a sortable table.

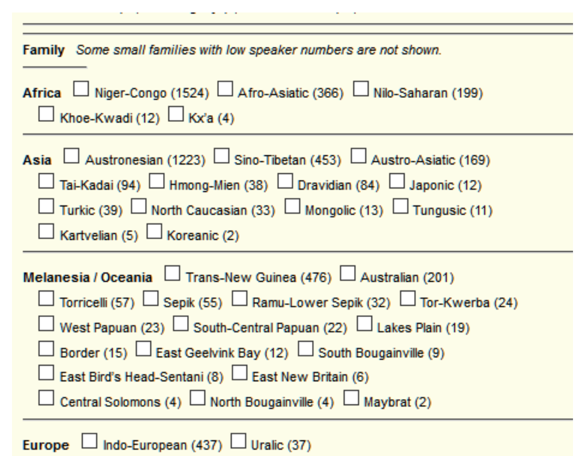


Below, language counts and national populations from Ethnologue 18. Countries that do not have "native" languages do not show language or population counts. ADM-1 counts are inferred from GADM 2.8.

Abb	country	ISO	ADM-1	Population
CHN	China	252	31	1,357,380,000
IND	India	381	37	1,252,140,000
USA	United States	181	52	313,914,000
IDN	Indonesia	684	34	248,818,000
BRA	Brazil	178	27	193,947,000
PAK	Pakistan	56	8	184,350,000
NGA	Nigeria	484	37	172,713,000
BGD	Bangladesh	15	7	156,591,000
RUS	Russia	91	85	143,856,000
JPN	Japan	14	47	127,339,000
MEX	Mexico	274	32	118,395,000
PHL	Philippines	175	81	98,394,000
ETH	Ethiopia	78	11	94,101,000
VNM	Vietnam	75	65	89,709,000

Prior to the query, the map area contained the list of country names, ISO 639-3 code counts, ADM-1 top-level administrative entities, and populations seen at left. This list has been re-sorted by clicking on the **Population** cell. While there are obvious exceptions, ADM-1 areas have reasonably consistent granularity





Query specification. The upper portion focuses on event types, while the lower portion allows specification of geographic areas and/or language families and branches (large portions of the center and bottom of the menu captures have been snipped).

These can be extended to refer to any aspect of the underlying data, and are implement as REST calls.

Center Top, after query "Myanmar"

7102 languages seen, **90** matched, in about 9 seconds. Showing event data for **1900-2016** only. Mouse over any column head for details, click to re-sort (shift+click for multiple sort columns). Click any **Language** or **Country** to see related events.

rank (%ile)	pop	ISO	language	region	country	ADM 1	national	regional	pivot	family	LDC	MT?	Glotto	orth	CRCL	events	dead	affected	$mil
37 (1%)	32035300	MYA	**Burmese**	SE Asia	**Myanmar**	Magway Region	mya T	mya T		Sino-Tibetan		G	**L**:D W C/**G**:G S/**P**:P	AC	HL Y1 Y3?	4	306	9,331,183	$177
191 (3%)	3295000	SHN	**Shan**	SE Asia	**Myanmar**	Shan State	mya T	shn	ETH: **lao** T **tha** TB	Tai-Kadai			**L**:D W C/**G**:G S	AC	HL Y1 Y3?	9	292	9,248,885	$122
269 (4%)	1800000	RHG	**Rohingya**	SE Asia	**Myanmar**	Rakhine State	mya T	rhg	GLO: **ben** TB	Indo-European					HL	12	609	9,696,525	$686
301 (5%)	1480000	KSW	**Karen, S'gaw**	SE Asia	**Myanmar**	Bago Region	mya T	ksw		Sino-Tibetan			**L**:D W C/**G**:G S/**P**:P	AC	HL Y1 Y3?	4	138,387	2,472,400	$4,000
371 (6%)	1050000	KJP	**Karen, Pwo Eastern**	SE Asia	**Myanmar**	Kayin State	mya T	kjp		Sino-Tibetan			**L**:W C/**G**:G S/**P**:P			7	138,513	11,677,700	$4,119
391 (6%)	1000000	rki	**Rakhine**	SE Asia	**Myanmar**	Rakhine State	mya T	rhg	*GLO: mya T*	Sino-Tibetan			**L**:W C/**G**:S/**P**:P			12	609	9,696,525	$686
408 (6%)	940000	KAC	**Jingpho**	SE Asia	**Myanmar**	Kachin State	mya T	kac		Sino-Tibetan			**L**:W C/**G**:G S/**P**:P	AC	HL Y1	8	226	9,295,943	$119
430 (7%)	851000	mnw	**Mon**	SE Asia	**Myanmar**	Kayin State	mya T	kjp	*GLO: khm T srb T vie TB*	Austro-Asiatic			**L**:D W C/**G**:G S/**P**:P		HL Y1 Y3?	7	138,513	11,677,700	$4,119
446 (7%)	805700	prk	**Wa, Parauk**	SE Asia	**Myanmar**	Shan State	mya T	shn	*ETH: khm T srb T vie TB*	Austro-Asiatic			**L**:D W C/**G**:G S/**P**:P/**T**:T			9	292	9,248,885	$122
539 (8%)	563960	ahk	**Akha**	SE Asia	**Myanmar**	Shan State	mya T	shn		Sino-Tibetan			**L**:D W C/**G**:G S/**P**:P	AC		9	292	9,248,885	$122
540 (8%)	560740	blk	**Pa'o**	SE Asia	**Myanmar**	Shan State	mya T	shn		Sino-Tibetan			**L**:W/**P**:P		HL	9	292	9,248,885	$122

Above, the response to a query **"Myanmar"**. Each row shows a single language. All columns are sortable, and support shift+click for secondary sort keys.

- **7102 seen, 90 matched** The total number of ISO 639-3 codes considered (7,102), and found in Myanmar (90).
- **rank** the relative position of this language among all 7,100 languages, sorted by speaker population. The (x%) gives its percentile ranking.
- **pop** speaker population, per Ethnologue 18
- **ISO, language** ISO 639-3 code and formal language name. The ISO that has the largest speaker population is shown in small caps to indicate that it is a good candidate to be a language of communication, and/or to provide a model for orthography. This cell is ***actionable***. When clicked, the lower center from shows details of all events that affected speakers of this language.
- **region, country** the world is divided into conventional regions: with numbers of languages, the top-level regions are *Africa (2,138), Americas (1,065), Europe (286)* and *Asia-Pacific (3,613)*. Each region is then subdivided; e.g. *Africa* into *Eastern*, *Western*, *Northern*, *Middle*, and *Southern*. Each sub-regain can then be specified by country.
  The **country** cell is ***actionable***. When clicked, the lower center from shows details of all events that affected all areas of this country.

- **ADM-1** the top-level administriative district associated with the language.
- **national, regional** these are ISO codes of the country's national language(s), and the nominal regional language – the highest-population language in the current ADM-1. A **T** indicates availability of machine translation technology, while **B** and **M** indication that "big" and "medium" amounts of other data (grammars, dictionaries, corpora) exist. For example, in Indonesia, the national language is Indonesian, but a minority language like Javanese or Sunda may be the language of education in a given province. A third, local language is often spoken at home.
- **pivot, family** a pivot language is the language that is most likely to be useful as an intermediate translation tool, assuming that it has resources. This cell lists the current language's immediate sisters (in roman) or cousins (in *italic*), per Ethnologue and/or Glottologue. The same **T B M** code shows resource availability. The **family** is the conventional name of the language phylum.
- **LDC, MT?, Glotto, orth, CRCL** all show resource availability. **LDC** and **CRCL** indicate data sets and delivery years; **HL** means that a *HA/DR* lexicon is available from CRCL. **MT** refers to **G**oogle, **B**ing, **Y**andex, or **G**oogle**Dev**elopment. The **Glotto** codes indicate a "best guess" as to the availability of basic print resources: **L**exical, **D**ictionary, **W**ordlist, **C**omparative, **G**rammar, (**F**ull or **S**ketch), **P**honology, or **T**ext. This helps distinguish between (somewhat) documented and (mostly) undocumented languages. Note that as a rule, none of these resources are in e-form. Finally, **Orth** indicates that an e-corpus sample is available via *An Crubadan*.
- **events, dead, affected, $mil** These cells summarize all events that have affected the current row's ADM-1 (not the current row's language). We assume that **affected** equals 10*dead if no value is give, but do not attempt to estimate costs. For the moment, we do not divide the effects of events over multiple ADM-1s, so there will be some overcounting.

Center Bottom, after query "Myanmar"

Click any **language** or **country** in the table, above, to see associated events. Initial analysis of returned results follows in the **two** separate tables below.

Event types by number of languages affected and number of events per language. (People per event type per language can't be calculated yet.)

Storm	85		Chinbon Chin [cnb] 4	Mro-Khimi Chin [cmr] 4	Sumtu Chin [csv] 4	Chak [ckh] 4	Rohingya [rhg] 4	Rakhine [rki] 4
Flood	82		Rawang [raw] 6	Rakhine [rki] 6	Mon [mnw] 6	Lashi [lsi] 6	Kachin [kac] 6	Daai Chin [dao] 6
Landslide	62		Thaiphum Chin [cth] 2	Tedim Chin [ctd] 2	Tawr Chin [tcp] 2	Mün Chin [mwq] 2	Rawngtu Chin [weu] 2	Khumi Chin [cnk] 2
Earthquake	45		Shwe Palaung [pll] 2	Shan [shn] 2	Intha [int] 2	Zayein Karen [kxk] 2	Yinchia [yin] 2	Danu [dnv] 2

The table below estimates **investment language** rank and benefit. Four distinct roles are considered: **national language** (which include both statutory and de facto languages), **regional language**, which is the most widely spoken language in a given ADM-1 area, **pivot sister**, which is the closest relative with sizeable technical or data resources, and **pivot cousin**, which can fill in for a missing sister.

Every language may have as many as four roles. All figures shown for each language reflect only the events for which it is counted as a national, regional, sister, or cousin language.

NB: Scroll the table up to the top of the frame, then click (or shift+click) column heads to sort.

pop	role	investment language	ISO	links	family	LDC	MT?	orth	CRCL	events	dead	affected	$mil
32,035,300	1 National	Burmese	mya	90	Sino-Tibetan		G	AC	HL Y1 Y3?	73	417,742	81,715,825	$14,084
344,000	2 Regional	Tedim Chin	ctd	20	Sino-Tibetan			AC		9	247	9,296,843	$119
3,295,000	2 Regional	Shan	shn	18	Tai-Kadai			AC	HL Y1 Y3?	9	292	9,248,885	$122
100,100	2 Regional	Tase Naga	nst	16	Sino-Tibetan				HL Y1	6	332	9,040,410	$121
1,800,000	2 Regional	Rohingya	rhg	8	Indo-European				HL	12	609	9,696,525	$686
940,000	2 Regional	Kachin	kac	7	Sino-Tibetan			AC	HL Y1	8	226	9,295,943	$119
1,050,000	2 Regional	Pwo Eastern Karen	kjp	6	Sino-Tibetan					7	138,513	11,677,700	$4,119
150,000	2 Regional	Western Kayah	kyu	6	Sino-Tibetan			AC		2	138,383	2,420,300	$4,000
32,035,300	2 Regional	Burmese	mya	3	Sino-Tibetan		G	AC	HL Y1 Y3?	4	306	9,331,183	$177

The initial center bottom response to the "Myanmar" query. There are two tables (one fixed, one sortable) above.

- **fixed table** this summarizes the number of major events that affected the search query area. For each event type, we also summarize the number of speaker communities affected. Varying terrain can cause these to vary greatly.
- **sortable table** this table estimates **investment language** rank and benefit; i.e. the language(s) for which it would be most useful to have advanced resources.
- **pop** the language speaker population, per Ethnologue.
- **role** worldwide, communities tend to be multilingual. The most common second languages tend to be either the **national** language of education, or a **regional** language of province-, state-, or island-wide communication (which may also be a language of education).

  For our purposes, a **pivot sister** language is the closed *etymologically related* language that has "substantial" resources, preferably machine translation. A **pivot cousin** is a step removed. As a practical matter the fact that a language is a sister or cousin does not necessarily mean that it will be close or comprehensible. We have suppressed some (but not all) of the artifacts that result from relying on standard linguistic subgrouping; this can be improved. Note that a single language (like

Burmese may fill multiple roles: it is a national and regional language, and is also etymologically close to some, but by no means all, of the Sino-Tibetan languages spoken in Myanmar.

- **investment language** as mentioned earlier, the national language is most likely to have good technology support, but is not necessarily the best pivot language for bootstrapping MT and similar tools. In effect, each row provides data that assists decision-making on whether an investment language should be pre-provision, and what it should be.
- **ISO, family** the ISO 639-3 code, and conventional language family name.
- **links** the number of languages for which the current language plays the stated role. For example, Thai is listed as the sister of four Tai-Kadai languages spoken in Myanmar.
- **LDC, MT, ortho, etc.** Summary totals of resources and events, as in the center top table.

Center Top (repeated from above)

7102 languages seen, **90** matched, in about 9 seconds. Showing event data for **1900-2016** only. Mouse over any column head for details, click to re-sort (shift+click for multiple sort columns). Click any **Language** or **Country** to see related events.

rank (%ile)	pop	ISO	language	region	country	ADM 1	national	regional	pivot	family	LDC	MT?	Glotto	orth	CRCL	events	dead	affected	$mil
37 (1%)	32035300	**MYA**	**Burmese**	SE Asia	**Myanmar**	Magway Region	mya T	mya T		Sino-Tibetan		G	**L:**D W C/**G:**G S/**P:**P	AC	HL Y1 Y3?	4	306	9,331,183	$177
191 (3%)	3295000	**SHN**	**Shan**	SE Asia	**Myanmar**	Shan State	mya T	shn	**ETH: lao T tha TB**	Tai-Kadai			**L:**D W C/**G:**G S	AC	HL Y1 Y3?	9	292	9,248,885	$122

Center Bottom (following click of country "Myanmar" in center top)

121 events found (66 EM-DAT, 55 GLIDE). **WHITE** seen and counted for at least one ADM-1 in the (possibly restricted) search above. **GREEN** seen, but no ADM-1 recognized or language counted. **BLUE** GLIDE data, not counted (ADM-1 is not specified).

ID	From	To	Country	abb	Location	Type	Subtype	Deaths	Affected	$000
G 2016-000088	2016-08-24	-	Myanmar	MMR	A powerful 6.8 magnitude earthquake struck central Myanmar Wednesday, killing at least three people and damaging some 60 pagodas in the famous ancient city of Bagan.	Earthquake				
G 2016-000092	2016-08-19	-	Myanmar	MMR	Tropical storm Dianmu formed in the South China Sea on 17 August and passed through Lao PDR around 2 days later, causing additional heavy rain which has been occurring since 11 August. Currently several districts in Luangprabang, Houaphan and Xaingabouli are affected, as indicated below.	Flash Flood				
G 2016-000058	2016-06-09	-	Myanmar	MMR	Heavy monsoon rains since the beginning of June have caused flooding in five states and regions of Myanmar. According to the initial reports from the Government Relief and Resettlement Department, at least 26,000 people are affected in Ayeyarwady, Bago and Sagaing regions as well as Chin and Rakhine states. A total of 14 deaths have been reported from the Union-level Relief and Resettlement Department, media sources and the Rakhine State Government.	Flood				
E 2016-0232	2016-06-01	2016-06-24	Myanmar	MMR	Sagaing, Bago regions, Rakhine state	Flood	--	14	3000	++
E 2016-0224	2016-05-23	2016-05-23	Myanmar	MMR	Hpakant region	Landslide	Landslide	42	15	++
G 2016-000052	2016-05-20	-	Myanmar	MMR	Tropical Cyclone ROANU continued moving north-east over the western Bay of Bengal, near the eastern coasts of India, retaining its intensity. On 20 May at 0.00 UTC its centre was located approx. 80 km south-east of Srikakulam district (Andhra Pradesh state, India) and it had max. sustained wind speed of 83 km/h. Over the next 48 h, the cyclone is forecast to strengthen as it continues moving north-east. It may reach Chittagong division (Bangladesh) on 21 May with estimated max. sustained winds of 100-130 km/h. Heavy rain, strong winds and storm surge are expected to affect southern Bangladesh and western Myanmar/Burma. A storm surge of 1.5 m is expected on the coastal area of Kutudbia (Cox's Bazar, Bangladesh) on 21 May morning (UTC).	Tropical Cyclone				
E 2016-0189	2016-04-29	2016-05-03	Myanmar	MMR	Mandalay city	Storm	Convective	18	87944	2600

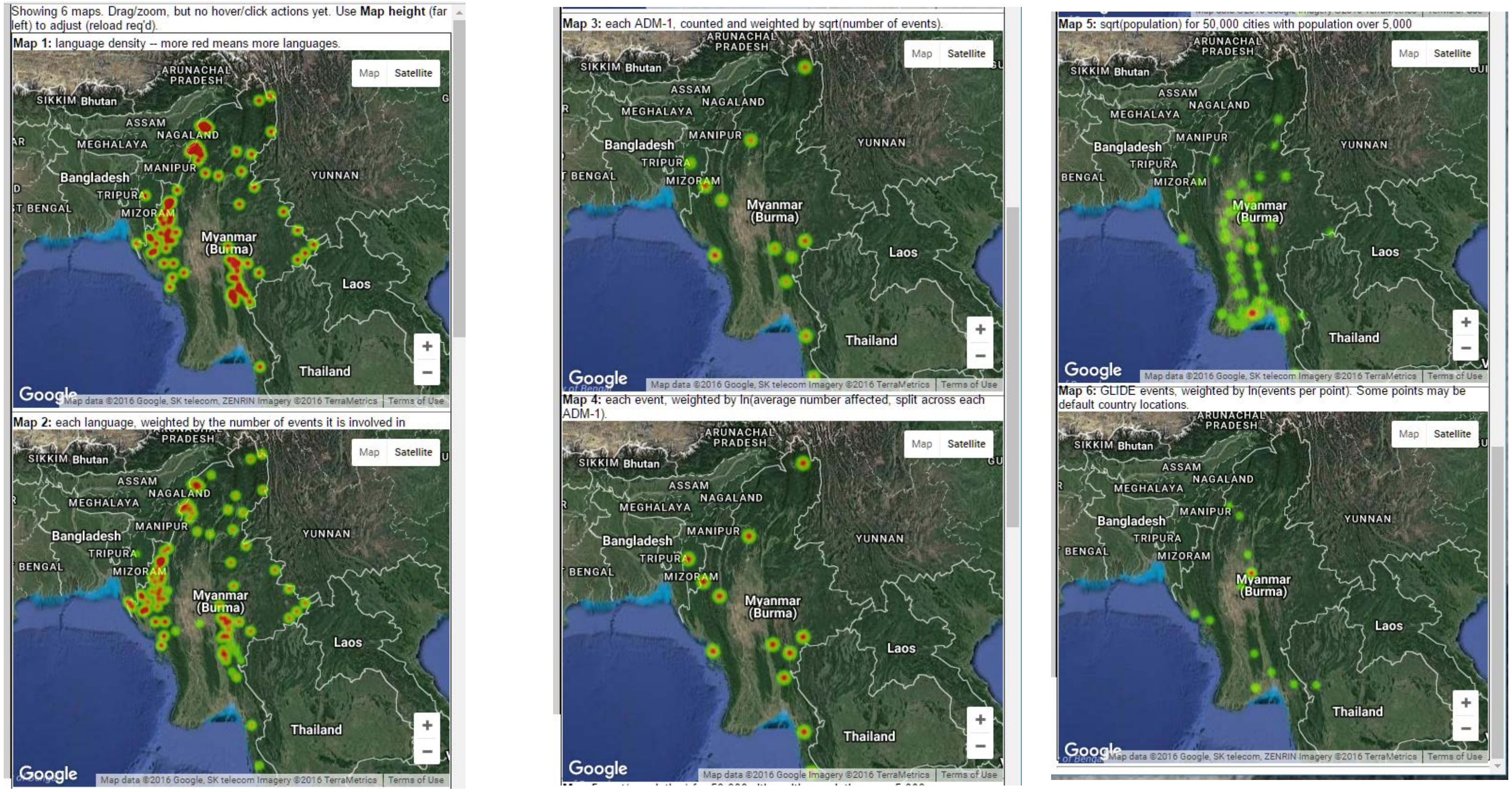
At present, clicking a **language** or **country** cell drills down to the related events. Above,121 events were reported for the "Myanmar" query: 66 from EM-DAT (given linked "E nnn" numbers), and 55 from the GLIDE set (given "G nnn" numbers). The different background colors are:

- **white** we were able to properly extract at least one ADM-1 area for this event from the EM-DAT dataset (which provides relatively regular listing of locations). The "E" number is actionable – in effect, it pre-populates the "Find a single EM-DAT#" text entry in the menu, then searches for all languages and ADM-1s associated with that event.
- **green** we were able to match the country (Myanmar), but not the ADM-1 entity (*Hpakant region* could not be parsed).
- **blue** GLIDE data. These reports have much more descriptive detail, but there is no regular encoding of casualties, costs, or impact area. We intend to align the GLIDE and EM-DAT datasets.

As noted, both EM-DAT and GLIDE numbers are used in other disaster-reporting contexts. We intend to:

- provide access to the raw EM-DAT and GLIDE data, and
- attempt to locate and link to any external data or sites related to the individual events.

Maps (far right)



We generate six heat maps based on the query for demonstration purposes (they render all 7,100 points very quickly). None of these maps are actionable, but that is an obvious next step. They are:

- **Map 1** language density in the query area (in this case, Myanmar).
- **Map 2** each language is weighted by the number of events it is involved in.
- **Map 3** each of Myanmar's 15 ADM-1 regions is treated as a centroid point, and weighted by number of events.
- **Map 4** each event is weighted by the log of the average number of people affected, split across affected ADM-1s.
- **Map 5** relative populations for all cities > 5,000.
- **Map 6** GLIDE events (which are given lat/long points), weighted by number of events/point (GLIDE sometimes uses a single point as the nominal location of many events).

This page is intentionally blank.

Appendix F: Tool snapshots

(taken from the project's Y1Q3 report)

Tool snapshots

CRCL is willing to provide access to many of our internal tools to other LORELEI performers. There are four web-based platforms:

~project/lorelei/data tools that focus on exploration of source texts. They provide highly detailed overviews and analyses of all data within one or more lects found within a given text.

~project/lorelei/dict tools that allow more traditional dictionary queries based on semantic and phonological criteria. Sources may be restricted by author, language, phylogenetic subgroup, or geographical region or proximity.

~project/lorelei/cogs the tools we use for exploring and creating cognate sets. They incorporate functionality for *semantic fallback* also see on the **/dict** page.

~project/lorelei/down the project download page. At this point we only link to prepared sets. However (given the complexity of the other pages) we will probably build in hooks to allow preparation and download of customized sets.

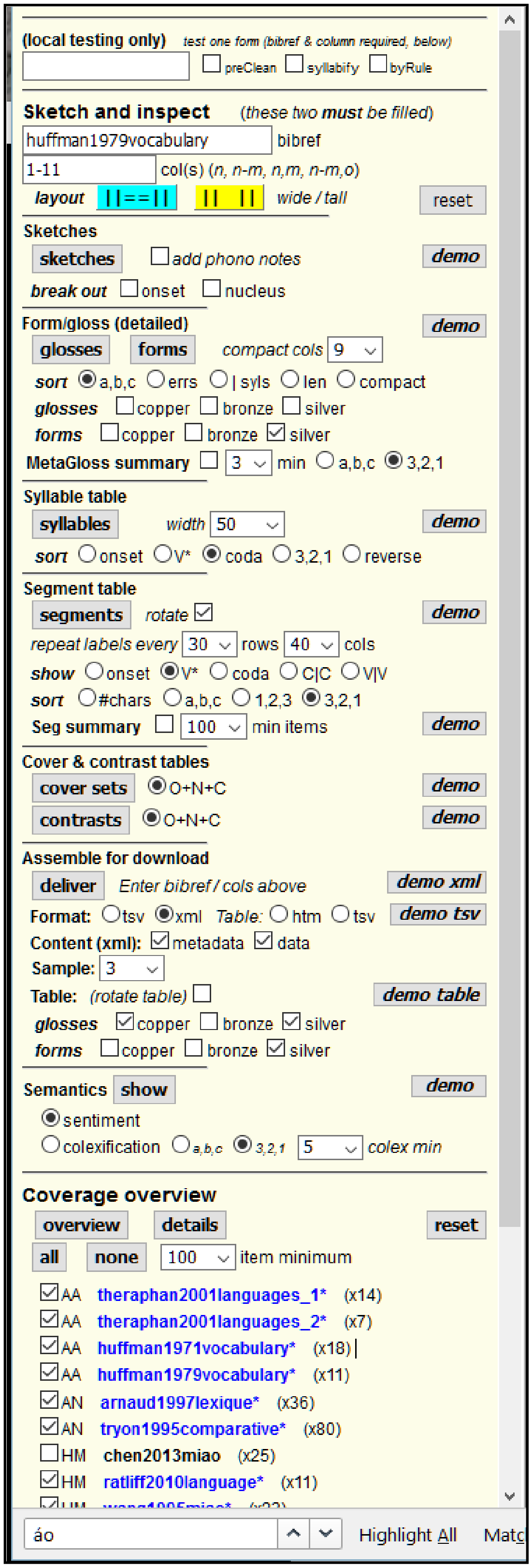
Please note that these pages are built by and for the CRCL development team. They are:

- beyond the scope of defined project deliverables, and not documented in detail,
- usually built to assist our own internal data audit and evaluation,
- subject to change at any time, and not guaranteed to be stable or persistent.

We are exposing them in order to:

- reveal the full extent of our datasets, including implicit as well as explicit content,
- clarify our capacities for data analysis and extraction,
- encourage requests for non-traditional data applications.

Essentially all functionality is provided by REST calls, and could be made accessible via external **http** queries (i.e. for machine-handling of returned data). Indeed, it must be understood that the purpose of many of these tools is simply to instantiate and help visualize (for testing purposes) the results of information extraction functionality that, in the long run, will be used in machine-to-machine communications.



/data overview

Normalizing transcribed data seems simple, but given many sources and ill-defined transcription systems (sometimes co-occurring in a single text) producing results that are consistent and accurate is extremely difficult. This page provides our main overviews. We begin with a quick overview of the menu.

Sketch and inspect name relevant source texts (*bibrefs*) and lects (logical columns).

Sketches provides various content inventories and counts, including phonemes, onset/nucleus/coda segments, canonical syllable shapes, and the like.

Form/gloss presents tables, usually in a compact form, of gloss and/or phonological form content. Like the next few functions, it is intended to provide a quick overview of the content of a typical 500-2,500 item lexicon, and is mainly used to oversee the automated processes that control semantic and phonoloigcal normalization.

MetaGloss summary tabulates and counts all normalized gloss forms by our extended part-of-speech definitions. .

Syllable table tabulates individual lect content by syllables, allowing sorts by onset, nucleus, coda, and count.

Segment table provides a global view of various syllable constituents for *all* datasets in a source. The different view and sort options help spotlight each of the underlying conversion decision processes.

Seg summary extracts and analyzes all syllable components from the complete dataset. It reveals the low-frequency elements that are more likely to be errors, and provides a basic sanity check on the dataset as a whole.

Cover & contrast tables answer two questions: what is the (probably) smallest subset of word that demonstrates all of a language's phonological features, and what is the complete set of words that demonstrates all positional contrasts (***hat*** vs ***cat*** contrast onset *h/c*).

Assemble for download packages the contents of these sources for inspection or download.

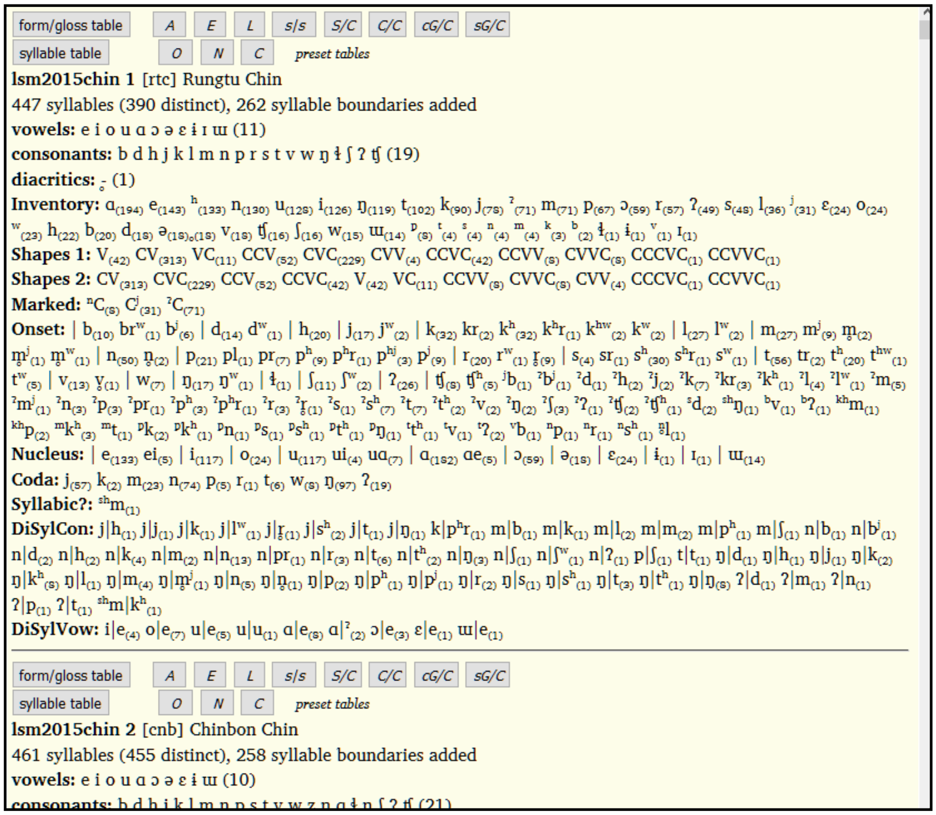
Semantics applies various measures of sentiment to lexicon semantics, and/or reveals co-lexification (use of the same word for different semantic concepts)

Coverage overview provides summary and detailed tables of linguistic coverage and content, excluding lects with fewer than some minimum number of items.

/data examples

Sketch As noted, our initial interest in this view is simply to get a bird's-eye view of the results of phonological conversion. There is a built-in mechanism (*add phono notes*) that displays any available data from PHOIBLE or the World Phonotactic Database. The **Shapes** are sorted first by length + alphabet, and then by frequency. The **DiSylCon** and **DiSylVow** entries show word-internal syllable boundary conditions for consonant and vowels. Many elements are *actionable* – a double-click on one of the shapes will find all forms with that shape.

As with other items, all suggestions regarding additions, refinements, and more convenient means of providing access to these data are welcome.



The **Shapes** are actionable, and trigger a source lookup. Below, all **CCVV** syllables; note that by design, the aspirated /pʰ/ is detected as a single character while the palatalized /pʲ/ forms are not:

Expansions: **C** to **"p|b|ɓ|m|ʙ|ɸ|β|w|ɱ|v|f|v|ʋ|θ|ð|t|d|ȶ|d̥|ɗ|n̥|n|r|ɾ|s|z|ɕ|ʑ|ʦ|ʣ|ʨ|ʥ|ɬ|ɮ|ɹ|ɭ̊|l|ʈ|ɖ|ɳ|ɽ|ʃ|ʒ|ʂ|ʐ|ɻ|ɭ|c|ɟ|**
V to "**i|y|ɨ|ʉ|ɯ|u|ɪ|ʏ|ʅ|ʮ|ʊ|ɥ|ʊ|e|ø|ɘ|ɵ|ɤ|o|ə|ɛ|ɛ|œ|ɜ|ɞ|ʌ|ɔ|æ|ɐ|a|ɶ|ᴀ|ɑ|ɒ**".

8 items found, 8 items returned in 2 seconds (note limit of 50 items per doculect)

copper	silver	gloss	ISO	language	family	bibref
pʰʲɑe.lɛ	pʰʲɑe|lɛ	butterfly	rtc	Rungtu Chin	ST	lsm2015chin 1
pʰʲɑe	pʰʲɑe	thigh	rtc	Rungtu Chin	ST	lsm2015chin 1
p.ruɑ	pruɑ	man	rtc	Rungtu Chin	ST	lsm2015chin 1
n.ruɑ	ⁿruɑ	woman	rtc	Rungtu Chin	ST	lsm2015chin 1
pʲɑe	pʲɑe	to give	rtc	Rungtu Chin	ST	lsm2015chin 1
ʔ.pʰu.pʲɑe	ˀpʰu|pʲɑe	to pay	rtc	Rungtu Chin	ST	lsm2015chin 1
p.kuɑ	ᵖkuɑ	nine (persons)	rtc	Rungtu Chin	ST	lsm2015chin 1
t.vɑ/pʲɑe.kʰɑt	ᵗvɑ / pʲɑe|kʰɑt	half (quantity)	rtc	Rungtu Chin	ST	lsm2015chin 1

Forms A view of raw copper and machine-processed silver forms from Tryon's Austronesian data. The silver columns show normalization of the transcription, and syllabification of individual forms. These views let us review large amounts of data quickly, identifying whether irregularities are due to our processing, or were found in the raw data.

Max **1280** forms found. The row numbers are arbitrary, and rows in different columns are not related. Each column is sorted by **alpha**. You can adjust the content using the main menu's **form/gloss table**

row	Column/lect 1 [tay] Atayal *copper*	*silver*	Column/lect 2 [tsu] Tsou *copper*	*silver*	Column/lect 3 [dru] Rukai *copper*	*silver*	Column/lect 4 [pwn] Paiwan *copper*	*silver*	Column/lect 5 [tao] Yami *copper*	*silver*
0	aķβux	aq|βux	a-fku-fkuŋu	af|kuf|ku|ŋu	aoḻo	ao|ḻo	a-uta	aw|ta	abtan	ab|tan
1	aupun	aw|pun	afs-a	af|ʃa	ababay	a|ba|baj	alʸak	a|lʸak	abto	ab|to
2	ayiβaw	a|ji|βaw	a-fʔori	af|ʔɔ|ɽi	ababay	a|ba|baj	alʸak	a|lʸak	abtək	ab|tək
3	ayiŋ	a|jiŋ	akʔi	ak|ʔi	ababiraw	a|ba|bi|raw	alʸak	a|lʸak	ai	aj
4	amiɬ	a|miɬ	akʔe-ŋica	ak|ʔɛ|ŋi|tsa	a-ba-bōŋo	a|ba|bo:|ŋo	alʸak	a|lʸak	al	aj
5	amiɬ	a|miɬ	a-mso	am|ʃɔ	abakə	a|ba|kə	alʸak	a|lʸak	ai no aləla	aj no a|lə|lə
6	aķih	a|qih	a-pta-ptaiŋi	ap|tap|tai|ŋi	abarə	a|ba|rə	alʸak	a|lʸak	ai no vaʁay	aj no va|ʁaj
7	aķih tāɬ-an	a|qih ta:|ɬan	a-tpiti-a	at|pi|tia	abarə	a|ba|rə	alʸak	a|lʸak	akmi kaḍai	ak|mi ka|ḍaj
8	aķih ɣiʔ	a|qih ɣiʔ	atvoxi	at|vɔ|xi	abo	a|bo	alʸis	a|lʸis	akpəʁən	ak|pə|ʁən
9	aķih ʔa ʔutux	a|qih ʔa ʔu|tux	au-pciro	aup|tsi|ɽɔ	a-daili	a|daj|li	alʸu	a|lʸu	aktokto	ak|tok|to
10	ašiɣaʔ	a|ʃi|ɣaʔ	au-tʔo-tʔou	aut|ʔɔt|ʔɔw	akamə	a|ka|mə	alʸu-alʸu	a|lʸua|lʸu	aktəbən	ak|tə|bən
11	a-ši-βaķ-i-βaķ-i	a|ʃi|βa|qi|βa|qi	avʔu	av|ʔu	akoaḍao	a|koa|ḍao	a-nəma	a|nə|ma	akḍotən	ak|ḍo|tən
12	aŋriʔ	aŋ|riʔ	a-xtosi	ax|tɔ|ʃi	alo-alo	a|loa|lo	asaw	a|saw	amyətaŋ	am|jə|taŋ
13	hayriŋ	haj|riŋ	a-ko-koru	a|kɔ|kɔ|ɽu	a-ləbə	a|lə|bə	aŋalʸ	a|ŋalʸ	amlavi	am|la|vi
14	hayɬaɣ	haj|ɬaɣ	amo	a|mɔ	ama	a|ma	aḷay	a|ḷaj	amlokolokoŋ	am|lo|ko|lo|koŋ
15	hamhum	ham|hum	amo	a|mɔ	ama	a|ma	aḷu	a|ḷu	amloloṣ	am|lo|loṣ
16	hanku	han|ku	amo	a|mɔ	ama	a|ma	bibi	bi|bi	amnoʁo	am|no|ʁo
17	haukuʔ	haw|kuʔ	amo	a|mɔ	ama	a|ma	buḷa-buḷay	bu|ḷa|bu|ḷaj	amvoyog	am|vo|jog
18	hawtiʔ	haw|tiʔ	amo-coni	a|mɔ|tsɔ|ni	ama	a|ma	bucaķ	bu|tsaq	anyoy	an|joj
19	ha-haβil	ha|ha|βil	a-tavri-si	a|ta|vɽi|ʃi	ama	a|ma	bəru-bəruŋ	bə|ru|bə|ruŋ	anŋət	an|ŋət
20	hahlpay	ha|hl|paj	a-trurunu	a|tɽu|ɽu|nu	amay	a|maj	caynan	caj|nan	aob	aob
21	hahipux	ha|hi|pux	a-xoi	a|xɔj	amanikino	a|ma|ni|ki|no	cay-saŋa-saŋas	caj|sa|ŋa|sa|ŋas	aob	aob
22	ha-hirhir	ha|hi|rʰir	a-ŋare	a|ŋa|ɽɛ	anəanə	a|naa|nə	cay-viḷi-viḷilʸ	caj|vi|ḷi|vi|ḷilʸ	apnərak	ap|nə|rak
23	hakriʔ	ha|kriʔ	a-ŋaco	a|ŋa|tsɔ	apiakaə-kanə	a|pia|kaə|ka|nə	cay-viḷi-viḷilʸ	caj|vi|ḷi|vi|ḷilʸ	appa	ap|pa

Glosses These may be extracted and viewed standalone in order to make is easier for LORELEI collaborators to understand the content and ordering of the various comparative and survey elicitation lists. Below, part of the LSM2015 list. Note that while most entries are given as WordNet *form#POS#sense-number*, we rely on extended forms for many kin terms ("Obro.fem" = "older brother of female"), and other terms that are widely lexicalized in Asia-Pacific. A small amount of inconsistency and uncertainty are expected for these silver glosses.

Total **441** distinct forms found (longest list 446). Compact sort (in 6 columns). Showing **gloss** forms only. Multiple lists will be merged in a single table

I#p#1	Obro.fem#k#1	Obro.mal#k#1	Osis.fem#k#1	Ybro.fem#k#1	Ybro.mal#k#1
Ysis.fem#k#1	Ysis.mal#k#1	afraid#a#1	all#a#1	angry#a#1	ant#n#1
arm#n#1	armpit#n#1	arrow#n#2	ascend#v#1	ashamed#a#1	ashes#n#1
back#n#1	bad#a#1	bald#a#2	bamboo#n#2	bamboo_shoot#n#1	banana#n#2
bark#n#1	bark#v#4	barking_deer#n#1	bathe#v#3	bear#n#1	beard#n#1
bee#n#1	beer#n#1	belly#n#1	bend#t#3?	betel_nut#n#1	big#a#1
bird#n#1	birdnest#n#1	bite#v#1	bitter#a#6	black#a#1	blanket#n#1
blind#a#1	blood#n#1	blow#v#1	blunt#a#1|blunt#a#2?	body_hair#n#1	boil#t#2
bone#n#1	bow#n#4	brain#n#1	branch#n#2	breathe#v#1	buffalo#n#4
burn#t#1	bury#v#2	butterfly#n#1	buy#v#1	calf#n#2	cane#n#2
cat#n#1	cheek#n#1	chicken#n#2	chin#n#1	choose#v#1	clothing#n#1
cloud#n#2	cockroach#n#1	cold#a#1:feeling	comb#n#1	come#v#1	cook#t#2
cooked_rice#n#0	cool#a#1:object	corn#n#1	correct#a#1	cough#v#1	count#v#1
cow#n#1	crawl#v#1	crest#n#5	crocodile#n#1	crossbow#n#1	cut#v#1
dance#v#1	dark#a#1	day#n#4	deaf#a#1	deep#a#3	descend#v#1
die#v#1	difficult#a#1	dig#v#1	dirty#a#1	do_not#x#1	dog#n#1
door#n#1	dream#v#2	drink#v#1	drum#n#1	drunk#a#1	dry#a#1
dry#t#1	dust#n#1	ear#n#1	earthworm#n#1	east#n#4	easy#a#1
eat#v#1	egg#n#2	eggplant#n#1	eight#n#1	elbow#n#1	elephant#n#1
elephant_tusk#n#0	enter#v#1	exchange#v#2	excrement#n#1	extinguish#v#2	eye#n#1
eyebrow#n#1	eyelid#n#1	face#n#1	fall#i#1	far#a#1	fast#a#1
fat#a#1	father#n#1	feather#n#1	few#a#1	field#n#1:dry	field#n#1:wet
fight#v#1	fingernail#n#1	fire#n#3|fire#n#7	firewood#n#1	fish#n#1	five#n#1
float#i#1	flow#v#2	flower#n#2	fly#i#1	fly#n#1	forehead#n#1
forget#v#2	four#n#1	free#v#1	friend#n#1	frog#n#1	fruit#v#2
full#a#1	garlic#n#1	ghost#n#3	ginger#n#3	give#v#1	go#v#3
go_out#v#1	gold#n#3	gong#n#1	good#a#1	grassland#n#1	green#a#1
grind#v#5	gums#n#2	half#a#1	hard#a#3	hate#v#1	he&she#p#1
head#n#1	head_hair#n#1	head_louse#n#1	hear#v#1	heart#n#2	heavy#a#1
heel#n#2	hide#v#2	hit#v#3	horn#n#2:buffalo	hot#a#1:feeling	hot#a#1:object

Metagloss summary This summarizes the entire dataset. Below **m** marks modals, **x** are unassigned, **j** are conjunctions, **r** are adverbs, etc. Classes may be assigned algorithmically, e.g. we can distinguish open and closed-class adverbs. The items are WordNet 3.0 senses, with extensions as needed. New classes (e.g. **p**ronouns or **k**in terms) are numbered beginning with **1**, while the standard **n, v, a, r** classes are numbered **0**. Note that because of wide variation in raw glosses, and corresponding difficulty in disambiguating senses, in some cases precise assignments will not be completely resolved until we are further along in cognate grouping.

3909 distinct glosses from **441061** items. Minimum cutoff is **3**.
This view ignores **:modifier** elements, and treats **i**(ntransitive) / **t**(ransitive) items as **v**(erbs).

m: 2/41 **forms**/items	**x**: 5/456 **forms**/items	**j**: 7/713 **forms**/items	**d**: 5/1259 **forms**/items	**q**: **12**/3489 **forms**/items	**p**: **13**/5101 **forms**/items
33 must#m#1	269 do_not#x#1	286 and#j#1	609 that#d#1	667 when#q#1	1114 you#p#1
8 should#m#1	132 to#x#0	129 if#j#1	480 this#d#1	505 where#q#1	969 we#p#1
	38 per#x#1	116 because#j#1	84 these#d#1	503 who#q#1	601 he#p#1
	10 of#x#1	105 or#j#1	47 those#d#1	498 what#q#1	587 I#p#1
	7 passive#x#1	38 until#j#1	39 this_one#d#1	435 how_many#q#1	577 they#p#1
		37 but#j#1		241 how#q#1	392 she#p#1
				203 how_much#q#1	250 it#p#1
				185 why#q#1	213 some#p#1
				174 which#q#1	147 others#p#1
				43 how_few#q#1	86 my#p#1
				34 from_where#q#1	77 his#p#1
					51 oneself#p#1
					37 your#p#1

r: 75/5616 **forms**/items	**k**: **147**/13094 **forms**/items	**a**: **484**/64197 **forms**/items	**v**: **1177**/131383 **forms**/items	**n**: **1982**/228641 **forms**/items
341 before#r#1	582 fat#k#1	766 hot#a#1	771 cut#v#1	1334 rice#n#1
267 no#r#3	582 son#k#1	752 true#a#1	693 fall#v#1	854 chicken#n#2
226 on#r#0	577 hus#k#1	690 quick#a#1	639 bite#v#1	769 person#n#1
222 thus#r#2	541 mot#k#1	685 near#a#1	635 know#v#1	695 sarong#n#1
186 below#r#1	529 wif#k#1	618 old#a#1	622 weep#v#1	692 year#n#1
182 in_front#r#1	493 Ison#k#1	617 full#a#1	612 smell#v#1	689 paddy#n#3
179 after#r#1	322 Osis.fem#k#1	605 fat#a#1	604 tie#v#1	688 tree#n#1
175 again#r#1	303 dau#k#1	602 cold#a#1	595 lift#v#1	676 knife#n#1
168 above#r#2	279 Idau#k#1	569 dry#a#1	588 return#v#1	667 louse#n#1
163 behind#r#1	241 Ybro#k#1	568 narrow#a#1	583 dig#v#1	649 leech#n#1
145 not#r#1	236 fat.par#k#1	563 far#a#1	583 steal#v#1	639 meat#n#1
141 up#r#1	227 mot.par#k#1	552 hungry#a#1	580 give#v#1	637 stone#n#1
140 down#r#1	218 chi.chi#k#1	552 wrong#a#1	577 fly#v#1	627 house#n#1
138 always#r#2	214 chi#k#1	550 bitter#a#6	576 love#v#3	621 night#n#1
122 from#r#0	200 Ysis.fem#k#1	542 good#a#1	571 throw#v#1	617 lightning#n#1
117 never#r#1	200 Ysis.mal#k#1	537 wide#a#1	570 bring#v#1	610 hair#n#1

Syllable table Segments (onset, nucleus, coda) can be viewed in the context of single languages (as below), or as large comparative tables. Below, vowel segments – and the syllables they appear in – from two Kra-Dai languages (from Hudak 2008):

-	xɔŋ$_{(1)}$											
ɔː	nɔːk$^{33}_{(1)}$	dɔːk$^{12}_{(1)}$	hɔːk$^{12}_{(1)}$	mɔːk$^{12}_{(1)}$	nɔːk$^{12}_{(1)}$	sɔːk$^{12}_{(1)}$	tɔːk$^{12}_{(2)}$	ŋɔːk$^{12}_{(1)}$	ʔɔːk$^{12}_{(1)}$	kɔːp$^{12}_{(1)}$	jɔːt$^{33}_{(1)}$	m
ɛ	mɛm$^{31}_{(1)}$	tɛm$^{12}_{(1)}$	kɛm$^{11}_{(1)}$	lɛm$^{55}_{(1)}$	lɛn$^{33}_{(1)}$	cɛn$^{12}_{(1)}$	kɛn$^{12}_{(1)}$	pɛn$^{11}_{(1)}$	kʷɛn$^{55}_{(1)}$	tɛn$^{55}_{(1)}$	vɛn$^{55}_{(1)}$	xɛ
ɛː	lɛːk$^{33}_{(1)}$	bɛːk$^{12}_{(1)}$	tɛːk$^{12}_{(1)}$	xɛːk$^{12}_{(1)}$	ʔɛːk$^{12}_{(1)}$	mɛːp$^{33}_{(1)}$	hɛːt$^{33}_{(1)}$	dɛːt$^{12}_{(1)}$	pɛːt$^{12}_{(1)}$	hɛː$^{21}_{(1)}$	lɛː$^{21}_{(1)}$	pɛ
ɤ	kɤj$^{21}_{(1)}$	nɤj$^{12}_{(1)}$	pɤj$^{12}_{(1)}$	mɤj$^{55}_{(2)}$	pɤj$^{55}_{(1)}$	xɤj$^{55}_{(1)}$	dɤk$^{55}_{(1)}$	ʔɤk$^{55}_{(1)}$	lɤm$^{55}_{(1)}$	hɤn$^{21}_{(2)}$	tʰɤn$^{12}_{(1)}$	d
ɤː	cɤːk$^{33}_{(1)}$	lɤːk$^{33}_{(1)}$	ŋɤːk$^{33}_{(1)}$	lɤːk$^{12}_{(1)}$	pɤːk$^{12}_{(1)}$	pʰɤːk$^{12}_{(2)}$	ŋɤːk$^{12}_{(1)}$	lɤːt$^{33}_{(1)}$	dɤːt$^{12}_{(1)}$	hɤː$^{21}_{(2)}$	mɤː$^{21}_{(1)}$	x
ɯ	lɯk$^{33}_{(1)}$	dɯk$^{55}_{(1)}$	sɯk$^{55}_{(1)}$	tʰɯk$^{55}_{(1)}$	lɯm$^{21}_{(1)}$	jɯm$^{55}_{(1)}$	fɯn$^{21}_{(1)}$	mɯn$^{21}_{(1)}$	sɯn$^{21}_{(1)}$	xɯn$^{21}_{(1)}$	ŋɯn$^{21}_{(1)}$	m
ɯː	xɯːp$^{33}_{(1)}$	mɯːt$^{33}_{(1)}$	jɯːt$^{12}_{(1)}$	hɯː$^{21}_{(1)}$	jɯː$^{21}_{(1)}$	mɯː$^{21}_{(1)}$	tɯː$^{21}_{(1)}$	cɯː$^{33}_{(1)}$	sɯː$^{33}_{(1)}$	mɯː$^{31}_{(1)}$	sɯː$^{31}_{(1)}$	h

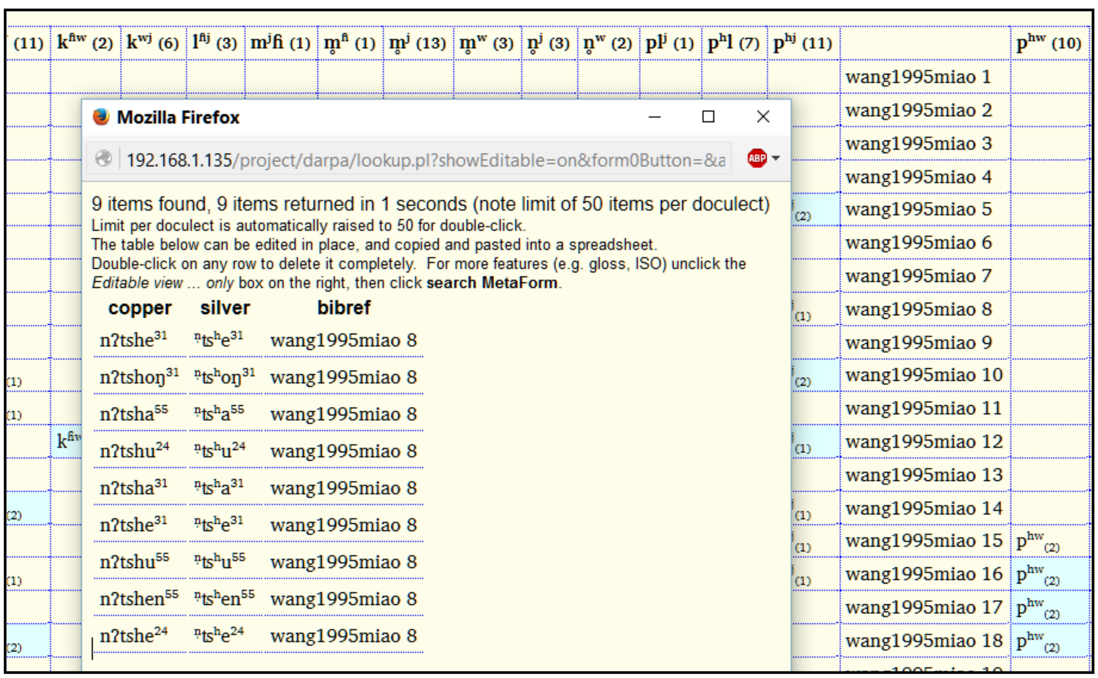
[khb] Lü. Sorted by nucleus, then onset. Syllabic (or stranded) C will appear first. You can adjust the content using the main menu's

a	pa$_{(1)}$	caj$^{33}_{(1)}$	haj$^{33}_{(1)}$	kaj$^{33}_{(1)}$	laj$^{33}_{(1)}$	naj$^{33}_{(1)}$	pʰaj$^{33}_{(1)}$	taj$^{33}_{(1)}$	xaj$^{33}_{(1)}$	caj$^{31}_{(1)}$	maj$^{31}_{(1)}$
-	pʰaj$^{45}_{(1)}$	saj$^{45}_{(1)}$	taj$^{45}_{(1)}$	tʰaj$^{45}_{(1)}$	xaj$^{45}_{(2)}$	hak$^{33}_{(1)}$	lak$^{33}_{(1)}$	mak$^{33}_{(1)}$	pʰak$^{33}_{(1)}$	sak$^{33}_{(1)}$	cak$^{45}_{(1)}$
-	han$^{45}_{(1)}$	kan$^{45}_{(1)}$	kʷan$^{45}_{(1)}$	man$^{45}_{(1)}$	pan$^{45}_{(1)}$	tan$^{45}_{(1)}$	xan$^{45}_{(1)}$	kap$^{33}_{(1)}$	lap$^{33}_{(1)}$	nap$^{33}_{(1)}$	cap$^{45}_{(1)}$
-	haw$^{45}_{(1)}$	kaw$^{45}_{(2)}$	pʰaw$^{45}_{(1)}$	saw$^{45}_{(1)}$	xaw$^{45}_{(2)}$	caŋ$^{33}_{(1)}$	naŋ$^{33}_{(1)}$	taŋ$^{33}_{(1)}$	xaŋ$^{33}_{(1)}$	jaŋ$^{31}_{(1)}$	caŋ$^{342}_{(1)}$
aː	ŋaːj$^{33}_{(1)}$	haːj$^{31}_{(1)}$	saːj$^{31}_{(1)}$	baːj$^{342}_{(1)}$	caːj$^{342}_{(2)}$	daːj$^{342}_{(1)}$	jaːj$^{342}_{(1)}$	kʷaːj$^{342}_{(1)}$	laːj$^{342}_{(2)}$	saːj$^{342}_{(1)}$	xaːj$^{342}_{(1)}$
-	xʷaːm$^{342}_{(1)}$	daːm$^{11}_{(1)}$	xaːm$^{11}_{(1)}$	haːm$^{45}_{(1)}$	laːm$^{45}_{(1)}$	naːm$^{45}_{(1)}$	saːm$^{45}_{(1)}$	tʰaːm$^{45}_{(1)}$	xaːm$^{45}_{(1)}$	xaːn$^{31}_{(1)}$	faːn$^{342}_{(1)}$
-	baːw$^{12}_{(1)}$	paːw$^{12}_{(1)}$	jaːw$^{11}_{(1)}$	haːw$^{45}_{(1)}$	kʷaːw$^{45}_{(1)}$	naːw$^{45}_{(1)}$	saːw$^{45}_{(1)}$	xaːw$^{45}_{(1)}$	haː$^{33}_{(1)}$	jaː$^{33}_{(1)}$	kaː$^{33}_{(1)}$
-	tʰaː$^{11}_{(1)}$	xaː$^{11}_{(2)}$	ʔaː$^{11}_{(1)}$	caːŋ$^{33}_{(1)}$	haːŋ$^{33}_{(1)}$	kaːŋ$^{33}_{(1)}$	laːŋ$^{33}_{(1)}$	xʷaːŋ$^{33}_{(1)}$	caːŋ$^{31}_{(1)}$	laːŋ$^{31}_{(1)}$	maːŋ$^{31}_{(1)}$
ă	că$_{(1)}$	xă$_{(1)}$									
e	dek$^{45}_{(1)}$	lek$^{45}_{(1)}$	sem$^{11}_{(1)}$	sem$^{45}_{(1)}$	ten$^{342}_{(1)}$	pen$^{12}_{(1)}$	sen$^{11}_{(1)}$	lep$^{33}_{(1)}$	cep$^{45}_{(1)}$	hep$^{45}_{(1)}$	kep$^{45}_{(1)}$
eː	veːk$^{33}_{(1)}$	jeːp$^{12}_{(1)}$	jeːt$^{12}_{(1)}$	xeːt$^{12}_{(2)}$	leː$^{342}_{(1)}$	meː$^{342}_{(1)}$	xeː$^{12}_{(1)}$				

Segment table We sometimes need to look at syllable components in order to understand their distribution (from a linguistic point of view), or as a more practical matter, to help explain apparent gaps in the source notation – differences between two lects may be real, or they might just be a consequence of the field worker's notation. Below, a sample from a complete set of onsets for all 23 Hmong-Mien languages in Wang 1995. These have been ordered longest-onset-first; other options include alphabetical order and frequency. The colored cells account for more than 5% of a given language's total:

	jʱ (6)	jʷ (5)	j̊ (9)	kl (65)	kʰ (59)	kʱ (2)	kʲ (119)	kʷ (65)	lʱ (14)	lʲ (29)	lʷ (8)	mʱ (12)	mʲ (29)	mʷ (27)	m̥ (64)	nʱ (6)	nʲ (9)	nʷ (5)	ŋ̊ (96)	pl (66)	pl̥ (3)	pʰ (87)	pʱ (9)	pʲ (103)
wang1995miao 1					kʰ(3)										m̥(6)				ŋ̊(5)			pʰ(7)		
wang1995miao 2								kʷ(4)					mʲ(3)		m̥(1)				ŋ̊(4)			pʰ(2)		pʲ(1)
wang1995miao 3					kʰ(3)										m̥(3)				ŋ̊(11)	pl(6)		pʰ(4)		
wang1995miao 4					kʰ(3)				lʱ(7)			mʱ(4)			m̥(3)	nʱ(1)			ŋ̊(9)			pʰ(3)		
wang1995miao 5					kʰ(3)								mʲ(3)		m̥(3)				ŋ̊(7)	pl(6)		pʰ(4)		pʲ(6)
wang1995miao 6					kʰ(4)										m̥(7)				ŋ̊(6)	pl(18)		pʰ(5)		
wang1995miao 7																				pl(2)	pl̥(3)			pʲ(1)
wang1995miao 8					kʰ(3)								mʲ(1)		m̥(4)				ŋ̊(7)	pl(5)		pʰ(4)		pʲ(6)
wang1995miao 9					kʰ(4)								mʲ(1)		m̥(6)				ŋ̊(5)	pl(4)		pʰ(6)		
wang1995miao 10					kʰ(3)		kʲ(9)	kʷ(4)					mʲ(1)		m̥(6)				ŋ̊(6)			pʰ(5)		pʲ(4)
wang1995miao 11			j̊(1)		kʰ(2)		kʲ(1)	kʷ(3)		lʲ(4)			mʲ(2)		m̥(3)				ŋ̊(4)			pʰ(6)		pʲ(2)
wang1995miao 12	jʱ(6)					kʱ(2)		kʷ(4)	lʱ(7)	lʲ(1)		mʱ(9)			m̥(3)	nʱ(5)			ŋ̊(5)			pʰ(4)	pʱ(9)	pʲ(4)
wang1995miao 13				kl(7)	kʰ(8)		kʲ(7)	kʷ(8)		lʲ(2)					m̥(3)			nʷ(2)	ŋ̊(5)	pl(2)		pʰ(4)		pʲ(4)
wang1995miao 14					kʰ(11)		kʲ(15)	kʷ(9)									nʲ(8)					pʰ(5)		pʲ(11)
wang1995miao 15		jʷ(1)			kʰ(4)			kʷ(1)		lʲ(4)	lʷ(1)		mʲ(5)	mʷ(5)	m̥(3)		nʲ(1)		ŋ̊(5)			pʰ(4)		pʲ(14)
wang1995miao 16		jʷ(1)			kʰ(3)		kʲ(5)	kʷ(3)		lʲ(4)	lʷ(1)		mʲ(4)	mʷ(5)	m̥(3)				ŋ̊(5)			pʰ(4)		pʲ(6)
wang1995miao 17		jʷ(2)		kl(7)	kʰ(2)		kʲ(7)	kʷ(6)			lʷ(1)		mʲ(1)	mʷ(5)	m̥(4)				ŋ̊(1)	pl(3)		pʰ(2)		pʲ(6)
wang1995miao 18		jʷ(1)	j̊(1)	kl(5)			kʲ(33)	kʷ(5)			lʷ(1)		mʲ(1)	mʷ(1)				nʷ(1)		pl(5)		pʰ(3)		pʲ(5)
wang1995miao 19							kʲ(16)	kʷ(4)					mʲ(1)							pl(1)				pʲ(10)
wang1995miao 20				kl(8)			kʲ(15)	kʷ(4)												pl(1)				pʲ(8)
wang1995miao 21			j̊(6)	kl(19)	kʰ(2)		kʲ(4)	kʷ(8)		lʲ(4)	lʷ(3)		mʲ(1)		m̥(6)				ŋ̊(5)	pl(6)		pʰ(6)		pʲ(5)
wang1995miao 22			j̊(1)	kl(19)	kʰ(4)			kʷ(8)		lʲ(8)	lʷ(1)		mʲ(2)	mʷ(2)				nʷ(2)		pl(10)		pʰ(3)		pʲ(5)
wang1995miao 23							kʲ(8)			lʲ(2)			mʲ(3)											pʲ(8)

These cells are also actionable; below, the /ˌⁿʦʰ/ onsets.



Seg summary Below, an overview of all onset, nucleus, coda, and tone sequences, sorted by frequency with the **50 min** option selected. Note that Zipf's Law holds – frequent items dominate. Ignoring very low frequency items deals with noise (which can usually be traced to errors in the original data), while having minimal impact on the size or representativeness of the full database.

These sequences are drawn from 863528 syllables from 444143 words in 558 doculects.
The presence of leading raised consontants reflects an attempt to stay true to raw sources, while providing a hint as to how they should be analyzed. In general, leading raised consonants would/could probably be lowered and treated as syllabic segments (in some cases with an unwritten epenthetic schwa, depending on the exact sequences). Some semi-analyzed data is included here, so some irregularity (such as apparent tone marks in vowel sequences) may be encountered.
Below, all columns are sorted using the specified **reverse** option.

Onset:	1734 distinct items,	805655 raw total,	294 distinct >= 50,	794050 total	98.5% of distinct items shown
Nucleus:	720 distinct items,	860330 raw total,	190 distinct >= 50,	856137 total	99.5% of distinct items shown
Coda:	330 distinct items,	328438 raw total,	54 distinct >= 50,	326843 total	99.5% of distinct items shown
On+coda:	1963 distinct items,	1134093 raw total,	309 distinct >= 50,	1121494 total	98.8% of distinct items shown
Tone:	95 distinct items,	253143 raw total,	52 distinct >= 50,	252810 total	99.8% of distinct items shown

rank	Onset	items	Nucleus	items	Coda	items	On + coda	items	Tone	items
1	k	65630	a	209363	ŋ	61298	n	94386	55	51435
2	t	65308	i	104752	n	48796	k	93841	33	35346
3	l	60476	u	94240	j	38381	t	85945	31	27134
4	m	55043	o	74691	ʔ	30258	m	82535	53	16543
5	p	46695	ɑ	62108	k	28211	ŋ	77297	4	14900
6	n	45590	e	57233	m	27492	l	67009	35	13350
7	s	38401	ə	52444	w	21379	j	58493	2	11333
8	r	32283	ɔ	24949	t	20637	p	57747	21	11134
9	b	28895	aː	23500	p	11052	ʔ	57347	13	9254
10	ʔ	27089	ɛ	18866	r	8803	s	42044	11	7204

For our own data audit purposes, sorting by the number of Unicode characters in the sequence is more useful, since longer sequences are more indicative of error, e.g. tone "2323" in the second row. The second onset, /**ⁿᵗʂʰ**/, is a subtler error – when /**n**/ was raised to indicate prenasalization, the /**tʂ**/ affricate was not properly recognized. The onset at rank 10 has an equivalent problem. This occurs because the IPA and Unicode do not treat all affricates in the same way. We fixed the whole class of errors with a minor code tweak that 'unifies' some two-character affricates that do not have pre-built digraphs.

rank	Onset	items	Nucleus	items	Coda	items	On + coda	items	Tone	items
1	ᵏkʰx	116	a̰ː	314	j̊	544	ᵏkʰx	116	11–21	173
2	ⁿᵗʂʰ	76	ḛa	308	j̰	440	ⁿᵗʂʰ	76	2323	419
3	tʂʰ	1157	ɔ̰ː	251	n̰	230	tʂʰ	1157	231	3174
4	kʰr	647	ṳː	250	ŋ̰	143	kʰr	647	214	915
5	pʰʲ	588	i̤a	198	m̰	111	pʰʲ	588	213	859
6	kʰʲ	559	e̤ː	182	gs	110	kʰʲ	559	132	610
7	pʰr	330	ṳa	167	jʔ	108	pʰr	330	454	602
8	kʰʷ	311	o̤a	156	lʔ	80	kʰʷ	311	343	446
9	ᵐbʷ	297	i̤ː	153	ᵐb	77	ᵐbʷ	297	314	380
10	ⁿᵈʐ	288	o̤ː	150	ᵑʔ	68	ⁿᵈʐ	288	112	325
11	ᵐtʰ	269	a̤ː	142	ᵊʔ	66	ᵐtʰ	269	323	240
12	kʰl	242	uã	130	ŋʔ	63	kʰl	242	312	239
13	ᵏkʰ	230	oːy	126	lʲ	62	ᵏkʰ	230	342	204
14	pʰl	201	ãː	118	ᵊŋ	59	pʰl	201	232	87
15	ⁿtʰ	188	ɛ̰ː	116	tⁿ	58	ⁿtʰ	188	545	84
16	ᵏtʰ	175	uɑ̰	103	pʷ	50	ᵏtʰ	175	212	50
17	ᵑkʰ	174	ṵː	96	ŋ	61298	ᵑkʰ	174	55	51435
18	ⁿtsʰ	173	iã	95	n	48796	ⁿtsʰ	173	33	35346
19	ⁿtr	173	ṳə	95	j	38381	ⁿtr	173	31	27134
20	ᵐkʰ	145	ə̰˞	91	ʔ	30258	ᵐkʰ	145	53	16543

Cover & contrast These tables help describe each language's internal variation, and also provided minimal datasets that are extremely useful for testing downstream applications. Below, we see the distinct onset, nucleus, and coda segments found in the complete dataset, followed by a list of 34 words that use them all in context. Because it is provided by a computationally feasible *greedy set cover* algorithm it is very likely (but not certain) to be the smallest such set.

777 words in list, 34 words required to cover **huffman1971vocabulary 3**

Tokens: onset x36, nucleus x33, coda x12

Onset:	b c cʰ d f h j j̊ k kl kr kʰ kʰʷ kʷ l m m̥ n n̥ p pl pr pʰ pʰʲ pʲ r r̀ s t tʰ w ŋ ɬ ɲ ʔ ˀn
Nucleus:	a ae ao aɔ aɛ a̤ a̤ɛ e ea e̤ e̤a i i̤ o oa oe o̤ o̤a u ua ṳ ɑ ɑe ɑɛ ɑː ɑ̤ ɔ ɔ̤ ə əɛ ə̤ ɛ ɛ̤
Coda:	c h j k l m n p t ŋ ɲ ʔ

saŋ|ʔoe	anus#n#1
tɑ̤|cɑp	arrive#v#1
hə|dua	at#r#0
bɑːŋ|kon	birth#t#1
ha|ret|mɔ̤t	blink#v#1
cʰim	blood#n#1
kʷi	cart#n#1
klə̤|ja̤t|pə|lɑʔ	clothing#n#1
ɬɑɛ	copper#n#1
kə̤p|tʰɑʔ	cover#v#1
kra̤ɛ	deer#n#1
ʔa|ca|ha|ʔuj	diviner#n#1
kə|ɲaŋ	dry_season#n#1
fɔh	fever#n#1
poa|wɑɲ	game#n#1
hə|naɛ|tɑp|ni̤h	grave#n#2
sac|ŋeak	green#a#1
nṳm|mo̤ŋ	have#v#1|exist#v#1
m̥ac	hook#v#1
te̤ak|pə|kʰak|lɑ|cə̤l	knot#n#2

Contrast Below minimal contrast sets for the same lect. They are unusual in contrasting full onset, nucleus, and coda segments. Hovering over anynumbered cell reveals the contrasting items. On the right, we see the contrast of various consonant codas with open vowel finals.

Contrast (**huffman1971vocabulary 3**): **nuclei**

No contrasts for: e̤ oe ua ɑe ɑː ɑ̤ ɛ̤

-	ae	aɔ	aɛ	a̤	a̤ɛ	e	ea	e̤a	i	i̤	o	oa	o̤	o̤a	u	ṳ	ɑ	ɑɛ	ɔ	ɔ̤	ə	əɛ	ə̤	ɛ	
0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0
a	1	1	2	3	2	1	4	4	6	1	6	3	3	1	3	1	8	2	8	-	14	-	4	3	a
ao	-	1	-	-	-	-	-	1	-	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
aɔ	-	-	1	-	-	1	-	-	1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
aɛ	-	-	-	-	1	-	-	-	-	-	-	-	1	1	-	-	-	2	-	-	-	-	1	1	aɛ
a̤	-	-	-	-	-	-	-	3	-	-	-	-	2	-	-	1	2	-	1	-	2	-	3	1	a̤
a̤ɛ	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	1	-	-	1	-	-	-	a̤ɛ
e	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	1	-	1	-	-	-	-	-	e
ea	-	-	-	-	-	-	-	2	1	1	2	-	1	-	-	-	2	-	1	-	-	-	-	-	ea
e̤a	-	-	-	-	-	-	-	-	1	-	1	1	1	1	-	1	2	-	4	-	4	1	2	2	e̤a
i	-	-	-	-	-	-	-	-	-	1	1	1	-	-	1	-	-	-	2	-	-	-	-	-	i
i̤	-	-	-	-	-	-	-	-	-	-	1	1	3	-	-	3	2	-	3	-	-	-	1	-	i̤
o	-	-	-	-	-	-	-	-	-	-	-	-	3	1	-	-	3	-	4	-	1	-	-	2	o
oa	-	-	-	-	-	-	-	-	-	-	-	-	-	1	1	1	1	1	1	-	1	-	3	-	oa
o̤	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	4	-	4	-	-	-	3	2	o̤
o̤a	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	o̤a
u	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-	3	2	u
ṳ	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	4	1	3	1	1	-	2	-	ṳ
ɑ	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	5	2	1	-	4	1	ɑ
ɑɛ	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	1	-	ɑɛ
ɔ	-	-	-	-	-	-	1	-	-	-	-	1	-	1	-	-	-	-	-	-	-	1	3	3	ɔ
ə	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	4	-	ə
ə̤	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	ə̤

i:a = ba/bi hacac/hacic hadac/hadic maʔ/miʔ ʔanaɛ/ʔinaɛ ʔanɔʔ/ʔinɔʔ

Contrast (**huffman1971vocabulary 3**): **codas**

No contrasts for: l

-	c	h	k	m	n	p	t	ŋ	ɲ	ʔ	
0	2	8	3	4	3	3	4	4	2	7	0
c	-	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
h	-	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
j	-	-	-	-	-	-	1	1	-	-	j
k	-	-	-	4	2	-	4	8	3	3	k
m	-	-	-	-	3	1	4	2	3	6	m
n	-	-	-	-	-	3	5	-	2	1	n
p	-	-	-	-	-	-	2	1	1	1	p
t	-	-	-	-	-	-	-	7	4	3	t
ŋ	-	-	-	-	-	-	-	-	4	5	ŋ
ɲ	-	-	-	-	-	-	-	-	-	4	ɲ

h:0 = da/dah kama/kamah kla/klah kɔ/kɔh lə̤/lə̤h pʰʲa/pʰʲah tɔ/tɔh ʔa/ʔah

Assemble for download This provides variations on **tsv, xml**, and **htm** views in which more or less metadata is provided. For example, this is an htm view of all 18 lects in this source (columns are glosses, rows are lects).

1971 Huffman, Franklin Unpublished vocabulary lists Huffman Papers, sealang.net/archives/huffman

Row	1	2	3	4	5	6	7	8	9	10
copper gloss	one	two	three	four	five	six	seven	eight	nine	ten
silver gloss	one#n#1	two#n#1	three#n#1	four#n#1	five#n#1	six#n#1	seven#n#1	eight#n#1	nine#n#1	ten#n#1
1 [khm] Central Khmer	muəj	piː	baj	buan	pram	pram|muəj	pram|piː	pram|baj	pram|buan	dɑp
2 [lcp] Western Lawa	tiʔ	lə|ʔa	lə|ʔue	paon	pʰɔn	lɛh	ʔa|lɛh	sleʔ	ˢtaem	
3 [mnw] Mon	mo̤a	ba	pɑɛʔ	pɔn	pa|sɑɲ	ka|raɔ	ha|pɔh	ha|cam	ha|cic	cɔh
4 [mnw] Mon	mṳa	ba	pɑɛʔ	pɔn	a|sɔɲ	a|raɔ	ha|pɔh	ha|cam	ha|cic	cɔh
5 [cbn] Nyahkur	mṳaj	baːr	piːʔ	pan	cʰuːn	tiaw	ᵐpɔh	ᶮcaːm	ᶮciːt	cəs
6 [mlf] Mal	mə|ʔan	piəw	pliɛʔ	pʰoːn						
7 [kjg] Khmu	moːj	baːr								
8 [kdt] Kuy	mṳːj	bia	paj	pɔːn	səːŋ	tə|pa̤t	tə|pʰo̤l	tə|kṳal	tə|kɛ̤h	ᶮcət
9 [kdt] Kuy	muaj	baːl	paj	poːn	səːŋ	tə|pʰat	tə|pʰoːl	tə|kʰɔːl	tə|kʰɛːh	mə|cit
10 [bru] Eastern Bru	muəj	baːr	paj	pʉːn	siːŋ	tə|pa̤t	tə|pṳːl	tə|kṳal	tə|ke̤ːh	mə|cit
11 [ngt] Ngeq	məːc	baːr	pɛː	puən	səːŋ	tʰə|pʰat	tʰə|pʰoːl	tə|kɔːl	tə|kiəs	mə|cʰit
12 [alk] Alak	moːj	pʰaːr	paj	poːn	pʰar|tʰam	ta|raw	təm|pɔh	ta|ŋ̊aːm	taŋ|ciːn	cʰit
13 [lbo] Laven	muːj	bəːr	pɛː	puan	sɔːŋ	traw	pɑh	tʰaːm	ciːn	cet
14 [brb] Lave	muːj	baːr	peː	puan	cʰəːŋ	traw	pəh	tʰaːm	ceːn	cit
15 [sti] Bulo Stieng	muaj	baːr	peː	puan	pram	praw	pɔh	pʰaːm	ʃen	ɟə|mʌt
16 [tpu] Tampuan	maoɲ	pʰiar	paeŋ	pʷan	pə|tam	trao	tɨm|paoh	tɨ|ŋ̊aːm	ᶮcʰin	ᶮcʰɨt
17 [cog] Chong	mṳːj	pa̤ːj	pʰe̤ːw	pʰo̤ːn	pʰram	tɑːŋ	nṳːj	tiː	ca̤ːj	ra̤ːj
18 [pcb] Pear	moj	pe̤aːk	pʰek	pʰaː|on	pʰraːm	krɔŋ	ᵏʰnuːl	krɔ|ti	kaːn|seər	raːj

Sets can be rotated in place for easier browsing. Below, columns are lects and rows are glosses:

1995 Tryon, Darrell T. (ed.) Comparative Austronesian Dictionary. An Introduction to Austronesian Studies Berlin, New York: De Gruyter Mouton.

Row	copper gloss	silver gloss	1 [tay] Atayal	2 [tsu] Tsou	3 [dru] Rukai	4 [pwn] Paiwan	5 [tao] Yami	6 [isd] Isnag
1	world	world#n#1				ka|ɟu|na|ŋan	ka|za̤|wan	ka|la|wa|gaːn
2	earth, land	land#n#4	rauq	ˣpɨx|pɨ|ŋɨ	daə|daə	qi|pu	ta|na	lu|saʔ
3	earth-ground, soil	soil#n#2	rauq	tsroa	daə	qu|na|vu|lʲan	ta|na	lu|saʔ
4	dust	dust#n#1	ʔa|βiŋ	ron|pu|xu	θo|vo|go	lʲi|tsaq	liŋ|bo	taː|puʔ
5	mud	mud#n#1	ʧɬaq	ɗiŋ|ki	i|ḽi|tsi	vu|das	ə|tək	lu|paŋ
6	sand	sand#n#1	βu|na|qij	fur|fu|ʔu	ə|naj	ga|du	a|naj	gi|nat
7	mountain, hill	mountain#n#1	ra|ɣi|jax	fur|ŋu	ḽə|gə|ḽə|gə	da|ŋa|da|ŋas	to|kon	ban|taj
8	cliff, precipice	cliff#n#1	ɬu|hij	tiʔ|nɨ	to|ka|ɖa|nə	qu|ma	a|la|ʂʲaʂ	ba|gi
9	plain, field	plain#n#1	quiʃ	ᵑre|saŋ|si	ɖa|ta|nə		ka|za̤|ta|jan	ir|ʔir|ʔer
10	valley	valley#n#1	u|quʔ				do ka|ʂo|pi|tan no to|kon	ta|naːp
11	island	island#n#1					ma|ʁa|taw	pu|gu
12	mainland	mainland#n#1					poŋ|ʂo	
13	shore	shore#n#1	ʃʲaɣ		ba|bʲa|bi|la	lʲi|vu	ka|na|na|jan	dap|pit
14	cave	cave#n#1	kuːɣ	fro|ŋo	ba|ro|ŋo|lo	za|lʲum	ar|ʧip	ab|but
15	water	water#n#1	qu|ʃi|jaʔ	ᵗˢxu|mu	a|tsi|laj	lʲa|vək	za̤|nom	da|num
16	sea	sea#n#1	βa|ruʔ	ti|pi	ba|jo	lʲa|vək	a|wa	be|baj
17	calm (of sea)	calm#a#2					ma|ʁa|naŋ	na|la|naj
18	rough (of sea)	roiling#a#1				bu|tsaq	mar|ɖa	nag|da|wal
19	foam	foam#n#1	βa|βut	fro|si	ḽa|po|tso		o|tab	buː|gaʔ
20	ocean	ocean#n#1	βa|ruʔ	tɨ|pɨ			a|wa	be|baj
21	lake	lake#n#1	wa|ʧi|ɬuŋ		ba|jo	ḽa|cuk	mi|bəb|nəŋ a za̤|nom	a|baj ja pi|suŋ
22	gulf, bay	bay#n#1					wa|wa	sul|bog
23	lagoon	lagoon#n#1					wa|wa	pi|suŋ
24	reef	reef#n#1					kəj|ʂa|kan	
25	headland, point	cape#n#1					pam|ʂan	pug|pu|gu
26	wave	wave#n#1	ni|na|waʃ	ˢmut|ɓuk|ɓu|ku|ru	bi|ka|bi|ki	ɖa|rulʲ	am|lo|ko|lo|koŋ	bal|nag
27	tide	tide#n#1						
28	lowtide	low_tide#n#1					mam|ʧi	
29	hightide	high_tide#n#1					mə|nəp	

The **xml** and **tsv** views are the basis of the project’s data distributions.:

:

```
<dataset id="huffman1971vocabulary.c1">
 <metadata>
  <reference>
   <id>huffman1971vocabulary</id>
   <doi>15144/huffman1971vocabulary</doi>
   <creator>Huffman, Franklin</creator>
   <title>Unpublished vocabulary lists</title>
   <date>1971</date>
   <publisher>Huffman Papers, sealang.net/archives/huffman</publisher>
   <lects>18</lects>
  </reference>
  <language>
   <languageCode scheme="iso639-3">khm</languageCode>
   <languageName scheme="iso639-3">Central Khmer</languageName>
   <latLong source="Ethnologue18">12.4671,104.5699</latLong>
   <latLong source="Glottolog2.6">12.0515,105.015</latLong>
   <country source="Ethnologue18">Cambodia</country>
   <country source="Glottolog2.6">Cambodia</country>
   <adm level="1" source="Ethnologue18">Kampong Chhnang</adm>
   <adm level="1" source="Glottolog2.6">Kampong Cham Province</adm>
   <population source="Ethnologue18">14224500</population>
  </language>
  <doculect>
   <id>huffman1971vocabulary.c1</id>
   <doi>15144/huffman1971vocabulary.c1</doi>
   <creator>CRCL</creator>
   <date>2015</date>
   <notation>IPA</notation>
   <analysis>broad</analysis>
   <forms>887</forms>
  </doculect>
 </metadata>
 <data>
  <item id="huffman1971vocabulary:C:c1.r1.gs1495.i2" iso639-3="khm" lang="Central Khmer">
   <forms>
    <form status="copper" analysis="broad" script="IPA">muəj</form>
    <form status="silver">muəj</form>
   </forms>
   <glosses>
    <gloss status="copper" lang="eng">one</gloss>
    <gloss status="bronze">one</gloss>
    <gloss status="silver">one#n#1</gloss>
   </glosses>
  </item>
  <item id="huffman1971vocabulary:C:c1.r2.gs1562.i19" iso639-3="khm" lang="Central Khmer">
   <forms>
    <form status="copper" analysis="broad" script="IPA">pii</form>
    <form status="silver">piː</form>
   </forms>
```


Colexification attempts to identify universals related to semantic shift and inherent polysemy and heterosemy. Below, we look at Tryon Austronesian (a collection of 80 lects, here numbered); identifying all semantic pairs that are expressed with the same word in multiple languages.

```
die#v#1 kill#v#1        11|12|12|14|15
difficult#a#1   hard#a#3        1|40|52|59|60|61|63|67|70|78
difficult#a#1   heavy#a#1       33|42|43|45|48|49|50
dinner#n#1      supper#n#1      1|5|20|21|22|23|29|41|70
disappear#v#1   lose#t#1        8|11|14|18|27|29|32|35|46|75|79|80
dish#n#2        plate#n#4       2|5|6|8|9|10|12|13|14|21|22|23|26|28|35|48|51|52|54|62|63|63|64|65|66|67
ditch#n#1       furrow#n#1      2|28|69|70|71|80
divide#t#1      separate#t#2    16|17|36|40|43|47|48|61|73|75|77|78
divide#t#1      share#v#4       11|14|15|17|18|19|20|21|23|27|28|30|32|36|37|39|41|52|56|59|60|61|62|64|66|69|71|75|80
do#v#1  work#n#1        8|12|13|14|36|70|80
donkey#n#2      mule#n#1        5|18|19|29|60
doorpost#n#1    pillar#n#5|pole#n#1     16|17|21|33|36|47|48|60|75|79
dormitory#n#2:male      meeting_house#n#0       41|43|45|48|49|59|60|61|63|64|78
down#r#1|below#r#1      low#a#2 7|14|16|40|44|55|61|62
down#r#1|below#r#1      under#a#1       8|9|11|12|13|14|16|18|19|20|21|22|24|29|30|31|32|33|34|35|36|43|45|49|50|51|57|62|63|71|76|77|78|80
dribble#v#4     drip#v#1        7|11|29|31|33|35|45|56|80
drink#n#3       river#n#1       42|43|44|44|51|59|59
drink#n#3       water#n#1       17|41|42|43|44|50|51|59|70|80
drop#t#1        fall#i#1        7|9|11|12|15|29|32|34|35|44|54|55
drop#t#1        release#t#1     33|40|46|46|61|62
drown#i#3       sink#i#4        19|20|24|33|35|36|38|39|40|47|51|54|61|62|63|66|80
duck#n#1        goose#n#1       1|39|42|56|62|75
dwell#v#3       remain#v#2      6|15|19|20|21|25|26|27|29|30|31|33|36|37|39|42|43|48|49|50|52|61|62|63|64|65|71|72|73|74|76|78|80
dwell#v#3       sit#v#1 16|17|32|34|39|40|43|45|46|47|48|57|76|78|80
dye#v#1 paint#n#1       17|24|40|43|60|62|63|77
dye#v#1 paint#v#2       16|17|61|67|77
early#a#1       quick#a#1       13|34|44|60|73|74|77|78|79
early#a#1       soon#r#1        5|7|72|75|77
earn#v#1        find#v#3        12|34|35|36|52
earn#v#1        get#v#1 16|34|35|41|41|44|57|61|73|76|79
easy#a#1        light#a#1       34|36|42|45|48|52|61
eat#v#1 food#n#1|food#n#2       11|65|68|71|80
eat#v#1 meal#n#2        11|19|41|51|64|65|73|74|76
edge#n#1        side#n#1        4|8|12|14|24|51|67|78
egg#n#2 testicles#n#1   1|28|34|58|63
empty#a#1       zero#n#2|nothing#n#1    1|19|25|35|38|43|49
end#n#1 end#n#2 15|20|21|32|37|50|51|52|60|62|78
end#n#2 finish#v#1      11|25|30|34|43|49|70
end#n#2 last#a#2        10|16|22|25|27|35|57
end#n#2 stop#v#2        43|49|52|63|69|71
```

Semantics A question that arose for LORELE applications was whether the small lexicons this project is based on would have relevant semantic content. This feature looks at various measures of sentiment as applied to the Tryon Austronesian list.

Sentiment for **tryon1995comparative** gloss list. Using first items, without attributes (so *young#a#1|soft#a#1 -> young#a#1*).
SentiWords: Guerini M., Gatti L. & Turchi M. "Sentiment Analysis: How to Derive Prior Polarities from SentiWordNet". In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP'13), pp 1259-1269. Seattle, Washington, USA. 2013. hlt.fbk.eu/technologies/sentiwords
SentiWordNet: Stefano Baccianella, Andrea Esuli, Fabrizio Sebastiani. "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining". In Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10) (May 2010) sentiwordnet.isti.cnr.it
Valence, Arousal, Dominance: Warriner, A.B., Kuperman, V. and Brysbaert, M. (2013). "Norms of valence, arousal, and dominance for 13,915 English lemmas". Behavior Research Methods, 45, pp 1191-1207 crr.ugent.be/archives/1003

SentiWordNet positive		SentiWordNet negative		SentiWords positive		SentiWords negative		Valence positive		Valence negative		Arousal positive		Dominance negative	
1	praise#n#1	0.875	deny#v#1	0.867	happy#a#1	-0.880	murder#n#1	8.47	happy#a#1	1.48	murder#n#1	7.74	gun#n#1	2.14	earthquake#n#1
0.875	happy#a#1	0.875	moan#v#1	0.750	love#v#3	-0.865	rape#n#3	8	love#v#3	1.54	rape#n#3	7.24	rape#n#3	2.47	rape#n#3
0.75	fragrant#a#1	0.875	fear#n#1	0.737	faithful#a#1	-0.832	die#v#1	7.95	faithful#a#1	1.67	die#v#1	7.24	snake#n#1	2.5	enemy#n#1
0.75	good#a#1	0.75	difficult#a#1	0.722	good#a#1	-0.797	kill#v#1	7.89	good#a#1	1.81	kill#v#1	7.05	attack#n#1	2.61	slave#n#1
0.75	healthy#a#1	0.75	dirty#a#1	0.722	smile#v#1	-0.765	prison#n#1	7.89	smile#v#1	1.94	prison#n#1	6.91	spider#n#1	2.65	drown#i#3
0.75	beautiful#a#1	0.75	forget#v#2	0.697	waterfall#n#1	-0.762	bury#v#2	7.86	rest#i#2	1.95	bury#v#2	6.9	die#v#1	2.84	sick#a#1
0.625	innocent#a#1	0.75	danger#n#3	0.695	kiss#v#1	-0.762	debt#n#2	7.81	cure#t#1	1.95	debt#n#2	6.86	money#n#1	2.86	deaf#a#1
0.625	wise#a#1	0.75	scold#v#1	0.692	sweet#a#1	-0.760	hate#v#1	7.79	waterfall#n#1	1.96	hate#v#1	6.81	rich#a#1	2.94	drunk#a#1
0.625	love#v#3	0.75	pity#n#1	0.690	healthy#a#1	-0.755	vomit#v#1	7.78	kiss#v#1	1.98	vomit#v#1	6.81	kill#v#1	3	steal#v#1
0.625	easy#a#1	0.75	wrong#a#1	0.687	peace#n#1	-0.750	attack#n#1	7.77	sweet#a#1	2	attack#n#1	6.76	earthquake#n#1	3.04	blind#a#1
0.625	faithful#a#1	0.75	stupid#a#1	0.682	give#v#1	-0.735	slave#n#1	7.76	healthy#a#1	2.06	slave#n#1	6.75	lightning#n#1	3.04	shiver#v#1
0.625	wages#n#1	0.75	weep#v#1	0.670	new#a#1	-0.725	greedy#a#1	7.75	peace#n#1	2.1	greedy#a#1	6.62	laugh#v#1	3.05	tax#n#1
0.5	true#a#1	0.75	hate#v#1	0.662	praise#n#1	-0.710	poison#n#1	7.73	give#v#1	2.16	poison#n#1	6.6	flame#n#1	3.06	doubt#n#1
0.5	tall#a#1	0.75	sick#a#1	0.660	springtime#n#1	-0.707	steal#v#1	7.68	new#a#1	2.17	steal#v#1	6.55	scar#n#1	3.09	famine#n#2
0.5	naked#a#1	0.75	short#a#3	0.660	spring#n#3	-0.695	enemy#n#1	7.65	praise#n#1	2.22	enemy#n#1	6.48	crocodile#n#1	3.11	low#a#2
0.5	hand#n#1	0.75	boil#n#1	0.652	beautiful#a#1	-0.692	war#n#1	7.64	springtime#n#1	2.23	war#n#1	6.45	scorpion#n#3	3.12	fever#n#1
0.5	time#n#3	0.75	raw#a#3	0.647	tree#n#1	-0.685	widower#n#1	7.64	spring#n#3	2.26	widower#n#1	6.42	adultery#n#1	3.14	thief#n#1
0.5	clear#a#1	0.75	stinking#a#2	0.647	victory#n#1	-0.685	arson#n#1	7.61	beautiful#a#1	2.26	arson#n#1	6.35	gold#n#3	3.16	envy#n#2
0.5	warm#a#1	0.75	cold#a#1	0.645	daytime#n#1	-0.680	widow#n#1	7.59	tree#n#1	2.28	widow#n#1	6.33	fight#v#1	3.17	owe#v#1
0.5	rescue#v#1	0.625	bad#a#1	0.640	laugh#v#1	-0.680	convict#v#1	7.59	victory#n#1	2.28	convict#v#1	6.29	shout#v#2	3.18	crooked#a#1
0.5	light#a#6	0.625	fault#n#1	0.637	play#v#1	-0.677	sick#a#1	7.58	daytime#n#1	2.29	sick#a#1	6.27	weapon#n#1	3.2	sneeze#v#1
0.5	expensive#a#1	0.625	no#r#3	0.630	female#a#3	-0.670	thief#n#1	7.56	laugh#v#1	2.32	thief#n#1	6.27	war#n#1	3.22	suspect#v#1
0.5	clever#a#2	0.625	never#r#1	0.630	food#n#1	-0.667	grief#n#1	7.55	play#v#1	2.33	drown#i#3	6.26	hell#n#4	3.23	widower#n#1
0.5	strong#a#1	0.625	freeman#n#1	0.630	female#a#1	-0.660	deception#n#1	7.52	female#a#3	2.33	grief#n#1	6.26	hate#v#1	3.23	corpse#n#1
0.5	teach#v#1	0.625	unclear#a#2	0.625	heaven#n#1	-0.655	anxiety#n#1	7.52	food#n#1	2.36	deception#n#1	6.24	murder#n#1	3.26	grief#n#1
0.5	poet#n#1	0.625	guilty#a#1	0.625	bathe#v#3	-0.652	suspect#v#1	7.52	female#a#1	2.38	anxiety#n#1	6.21	impregnate#v#4	3.27	war#n#1
0.5	clean#a#1	0.625	mistake#n#1	0.625	sing#v#2	-0.652	lie#v#5	7.5	heaven#n#1	2.39	suspect#v#1	6.2	lick#v#2	3.28	strike#v#1
0.444	demon#n#1	0.625	unripe#a#1	0.625	warm#a#1	-0.650	grave#n#2	7.5	bathe#v#3	2.39	lie#v#5	6.14	fear#n#1	3.28	die#v#1
0.375	old#a#1	0.625	grief#n#1	0.625	summer#n#1	-0.642	rotten#a#2	7.5	sing#v#2	2.4	grave#n#2	6.1	expensive#a#1	3.29	threaten#v#2
0.375	similar#a#1	0.625	deception#n#1	0.620	silver#n#1	-0.637	corpse#n#1	7.5	warm#a#1	2.42	choke#t#4	6.1	famine#n#2	3.3	captive#n#1
0.375	proud#a#1	0.556	demon#n#1	0.620	hope#v#1	-0.635	cockroach#n#1	7.5	summer#n#1	2.43	rotten#a#2	6.05	magic#n#1	3.32	fear#n#1
0.375	anger#n#1	0.5	bad_luck#n#3	0.617	star#n#3	-0.632	ugly#a#1	7.48	silver#n#1	2.45	corpse#n#1	6.05	kiss#v#1	3.33	blame#n#1
0.375	embrace#v#2	0.5	sour#a#2	0.617	easy#a#1	-0.627	divorce#n#1	7.48	hope#v#1	2.45	hurt#i#1	6.05	fire#n#3	3.33	adultery#n#1
0.375	calm#a#2	0.5	certain#a#2	0.607	dream#v#2	-0.625	anger#n#1	7.47	star#n#3	2.46	cockroach#n#1	6.05	threaten#v#2	3.37	rotten#a#2
0.375	loud#a#1	0.5	blame#n#1	0.605	wise#a#1	-0.620	betray#v#2	7.47	easy#a#1	2.47	ugly#a#1	6.05	swim#v#1	3.38	murder#n#1

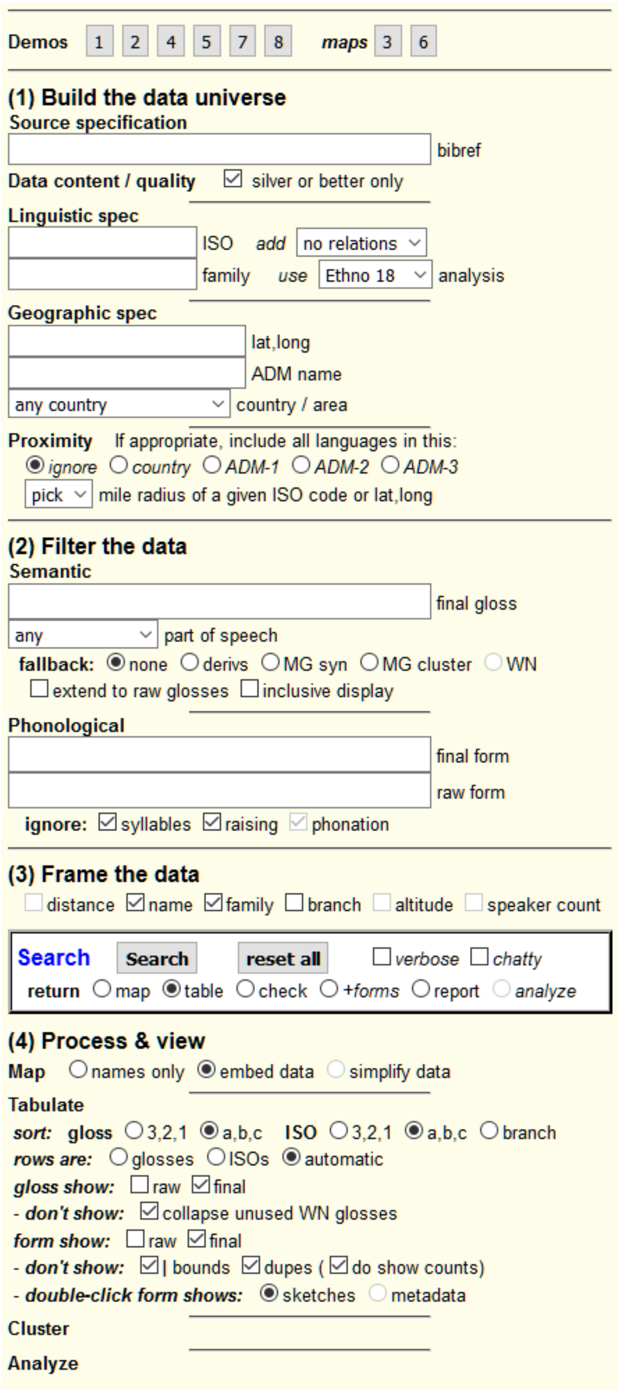
Coverage overview We saw the summary overview on page 1 of this report. The detailed view first summarizes all ISO codes, then lists sources with other details one by one

Sets per ISO code (larger numbers imply language surveys)

NONE:1	ace:1	acn:2	adi:1	adx:2	adz:1	aji:1	akl:1	alk:2	ane:1
anl:7	app?:1	atb:1	atq:2	ban:1	bbc:1	bca:1	bhz:1	bje:2	bkz:1
blt:1	bmt:2	bod:2	bpn:2	bps:1	brb:1	bru:2	brv:1	bug:2	bwx:2
bxd:1	bzh?:1	cam:1	cbn:1	cek:10	cgc:1	ckn:2	clj:3	clk:1	clt:8
cmr:10	cmw:2	cmy:1	cnb:9	cng:1	cnh:1	cog:1	cqd:1	csh:19	csv:8
cth:1	czt:1	dad:1	dao:26	ddg:1	dis:1	dru:1	dup:1	duu:1	enu:1
ero:1	ers:1	fij:1	gil:1	gor:2	gqi:1	hea:2	hlt:8	hmd:1	hmi:1
hmj:1	hml:2	hmm:2	hnn:1	how:1	huj:1	iii:2	ind:2	irh:1	irr:1
isd:1	ium:4	jae:1	jav:1	jeh:1	jiu:1	jmn:2	jya:1	kac:1	kaf:1
kdt:7	kem:1	kgc:1	kgd:1	khb:2	khm:1	kij:1	kix:1	kjc:1	kjg:1
kli:1	kmk:1	ksd:1	ksw:1	ktv:1	kuf:3	kvo:1	kwd:1	kzf:1	lbo:3
lcp:1	lew:1	lhu:1	lid:1	lis:1	llu:1	lpn:1	lsi:1	lus:1	lww:1
lzn:7	mad:1	mah:1	mak:1	mdh:1	mdr:1	mek:1	meu:1	mhs:1	mhu:1
mhx:1	min:1	mji:3	mjw:1	mkz:1	mlf:1	mmr:2	mna:1	mni:2	mnw:2
mqj:1	mqy:1	mrh:1	mrn:1	mro:3	mva:1	mvm:1	mvp:1	mw:1	mwq:5
mwt:1	mww:1	mxe:1	mxj:1	mya:1	nbe:1	nbi:2	nbu:2	nem:1	nen:1
ngt:4	njb:1	njh:1	njm:2	njn:1	njo:3	nkh:1	nki:1	nlq:1	nme:1
nmf:1	nmy:1	nng:1	nnl:1	nnp:1	nod:1	npg:3	nph:1	npo:1	npy:1
nqq:2	nqy:1	nre:1	nri:1	nsa:1	nsm:1	nst:51	ntx:3	nuf:1	nun:1
nut:1	nxa:1	nxg:1	nxk:2	nxq:1	nzm:1	oog:1	ors:1	pac:1	pcb:1
pha:2	plt:1	plw:2	pma:1	pmf:1	pmi:1	pmj:1	pnu:2	pon:1	ppk:1
pss:1	psw:1	ptt:1	pwm:1	pwn:1	pyu:1	pzn:4	qvy:1	rap:1	rog:1
rtc:4	rtm:1	rug:1	sas:1	sda:1	sez:8	shn:2	shx:1	skb:1	ski:1
smo:1	ssb:2	sse:1	sss:4	sti:1	sun:1	sxg:1	szw:1	tah:1	tao:2
tay:1	tbc:1	tbo:1	tdf:1	tet:1	tgl:1	tha:1	tih:1	tji:1	tnk:1
tnn:1	ton:1	tpu:1	tsj:1	tsu:1	tth:1	tto:2	tts:1	twh:1	twm:1
twu:1	umn:4	weu:4	wew:1	wlo:1	woe:1	wtw:1	wyy:1	xct:2	ycl:1
yim:1	ysn:1	ywq:1	zch:1	zeh:3	zgb:6	zgn:3	zha:7	zhb:1	zhd:1
zhn:2	zlj:2	zln:1	zqe:1	zyb:4	zyg:2	zyj:1	zyn:5	zyp:1	zzj:4

Details by bibref

ISO	items	bibref	col	lang
NONE	680	theraphan2001languages_2	6	Lavi
ace	1217	tryon1995comparative	17	Acehnese
acn	1471	huang1992tbl	29	Achang
acn	1676	huang1992tbl	28	Achang
adi	1770	huang1992tbl	24	Adi
adx	1345	huang1992tbl	5	Amdo Tibetan
adx	1702	huang1992tbl	4	Amdo Tibetan
adz	984	tryon1995comparative	49	Adzera
aji	1276	tryon1995comparative	66	A'jië
akl	1261	tryon1995comparative	9	Aklanon
alk	585	theraphan2001languages_2	4	Alak
alk	692	huffman1971vocabulary	12	Alak
ane	1069	tryon1995comparative	67	Xârâcùù
anl	451	lsm2015chin	122	Anu-Hkongso Chin
anl	451	lsm2015chin	123	Anu-Hkongso Chin
anl	452	lsm2015chin	118	Anu-Hkongso Chin
anl	454	lsm2015chin	119	Anu-Hkongso Chin
anl	454	lsm2015chin	124	Anu-Hkongso Chin
anl	456	lsm2015chin	120	Anu-Hkongso Chin
anl	458	lsm2015chin	121	Anu-Hkongso Chin
app?	670	tryon1995comparative	58	Raga
atb	1843	huang1992tbl	30	Zaiwa
atq	626	arnaud1997lexique	16	Aralle-Tabulahan
atq	663	arnaud1997lexique	15	Aralle-Tabulahan
ban	1190	tryon1995comparative	24	Balinese
bbc	1199	tryon1995comparative	18	Toba Batak
bca	1966	huang1992tbl	48	Central Bai
bhz	800	arnaud1997lexique	9	Bada (Indonesia)
bje	358	wang1995miao	22	Biao-Jiao Mien
bje	397	wang1995miao	21	Biao-Jiao Mien
bkz	901	arnaud1997lexique	29	Bungku
blt	968	hudak2008comparative	3	Tai Dam
bmt	400	ratliff2010language	10	Biao Mon
bmt	434	wang1995miao	18	Biao Mon
bod	1900	huang1992tbl	3	Tibetan



/dict

This page begins to develop the underlying functionality that will be required by more conventional dictionary applications. It takes an unconventional approach that is necessitated partly by the very, very large amount of date we provide access to, and partly by our anticipation of LORELEI's specific needs – in particular, the ability to focus or extend queries by region and relations.

(1) Build the data universe In effect, this step instantiates the dataset we wish to query. By default, queries are limited to silver-grade normalized datasets.

Linguistic spec define the universe in terms of ISO 639-3 codes, language family names (e.g. AA/AN/HM/KD/ST), or their analyzed phylogenetic relations. in the language family tree. Analyses vary; we support Ethnologue, Glottolog, and some local subgroupings.

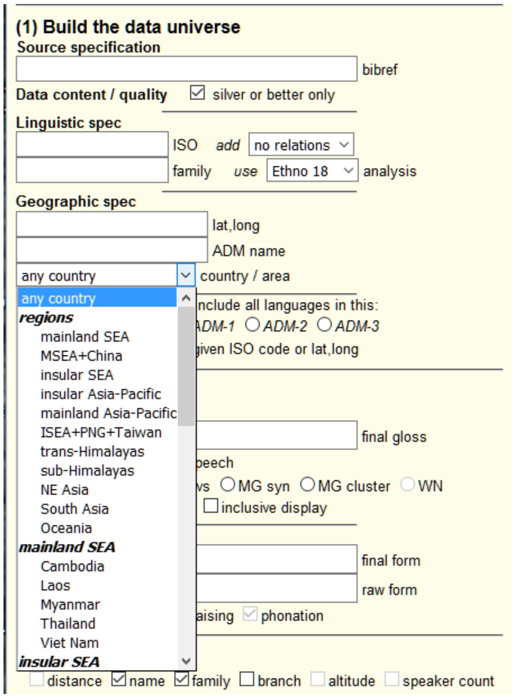
Geographic spec provide some means of defining, limiting, or extending a search. This is very helpful in regions with high language densities and mutual influence.

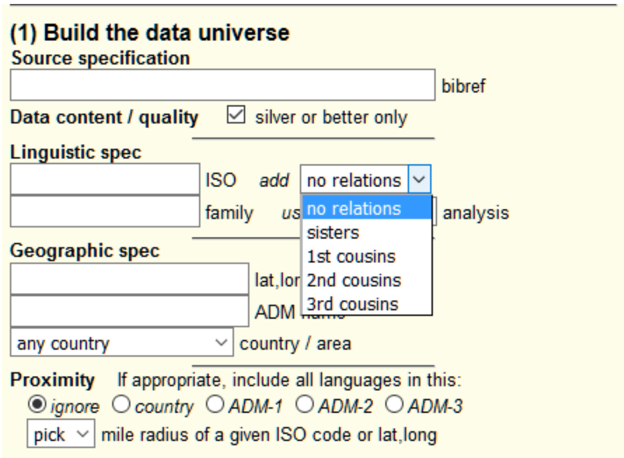
(2) Filter the data These provide what is ordinarily the semantic or phonological query. We are currently focused on facilities for *semantic fallback*; these are demonstrated below. The phonological search facility is limited at present.

(3) Frame the data Most of our knowledge about languages is actually external to the original data sources. Framing lets us add lect-specific facts to the returned forms and glosses, typically to aid in downstream applications (e.g. projection onto a map).

(4) Process & view Returned data will vary dramatically in size (from one item to thousands) and intended function. Beyond obvious alternatives of map or tabular views, we may wish to pass results to downstream applications (like our own apps in **/cogs**, discussed below). Again, we stress that these tools are not intended to produce a user-facing dictionary, but rather to help us instantiate and visualize this low-level functionality.

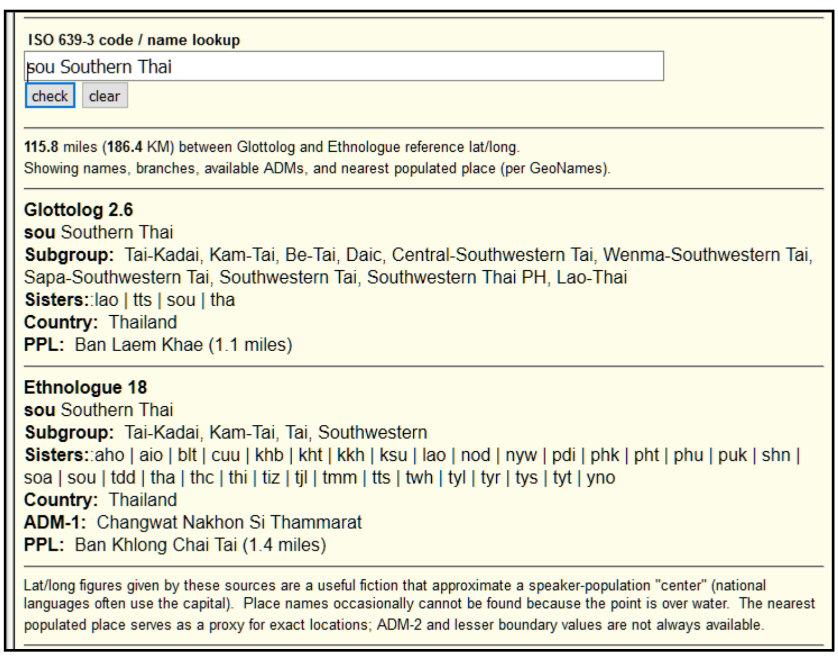
Build the data universe We can limit or extend the search universe by sources, phylogenetic linguistic specification, or geographical bounds / regions. This is important in areas for which data is limited because it lets queries fall back to languages that are related, or which are likely to be loan sources. Below, we show associated dropdown lists.





Now, the ISO 639-3 standard only specifies language names and three-letter codes. Information regarding phylogenetic subgrouping and speaker location must be provided by an external analysis. We track both Glottolog and Ethnologue, the only wide-scale analyses available.

The graphic below is produced by the **reference tools** widget on the far right of the **/dict** page; it shows the functionality underlying the **data universe** specification. The user enters the first few letters of a language code or name; we identify the proper code, then show geographic and subgrouping data as available. Note that these are by no means always in agreement – analyses and even locations may vary considerably.

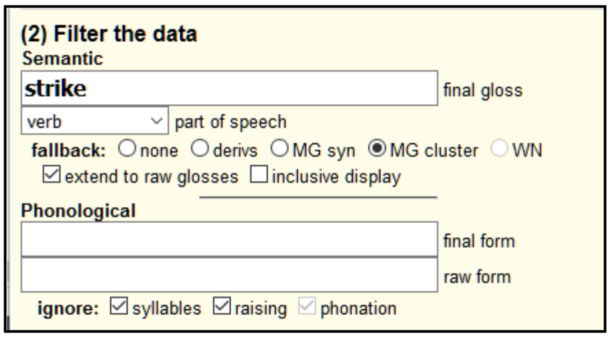


Because LORELEI-related responders are likely to be working with local civil authorities, we have gone to some lengths to attempt to identify speaker neighborhoods and enclosing regions in terms of formal ADM identifiers (and vice versa).

Filter the data This is what we ordinarily think of as formulating the query. Below left, we query **strike**. Part-of-speech can be specified, either to restrict a word sense, or to serve as a filter in place of any particular gloss. (e.g. we might request all kin terms).

Fallback controls semantic expansion. At present, options include *derivs* (English derived forms, e.g. “striking”, “striker”), the *MetaGloss* synonym set or cluster (semantically equivalent or related terms), or strict WordNet synonym sets. The *extend to raw glosses* option looks for the (possibly expanded) search term in the raw, copper gloss form as well as the normalized silver (or gold-standard) form. One consequence of expanding semantic targets is that a single lect may have multiple hits. Normally, we suppress secondary items – if the initial search form is found, expanded items are suppressed. *Inclusive display* returns all items all items.

As noted, phonological query options are limited at this point; available options include the ability to ignore syllable boundaries, and to treat raised items (which usually represent features or secondary phonemes) as though they were ordinary letters.



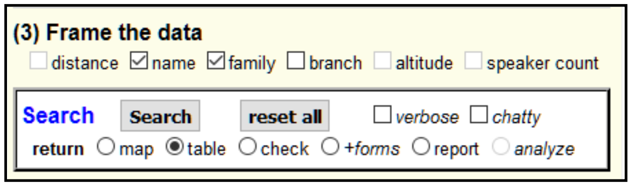
The result of this search is shown below.

Limiting datasets to silver or better (uncheck *silver or better* for higher/broader test volume):
huffman1971vocabulary|huffman1979vocabulary|theraphan2001languages_1|theraphan2001languages_2|hudak2008comparative|zhang1999zhuang|huan
Seeking gloss **strike**. Will fall back to raw glosses for any unmatched ISO slot.
Members of branch **KD** are:
aho|aih|aio|aou|blt|byk|cdy|cov|cuq|cuu|doc|enc|giq|gir|giu|giw|gqu|jio|khb|kht|kkh|kmc|ksu|kyp|lao|laq|lbc|lbt|lha|lic|lwh|mkg|mlc|mlm|mmd|nod|
Falling back to MetaGloss cluster **snakebite#n#1|v@snakebite#n#1|strike#v#0|hit#v#3|strike#v#1|deal_a_blow#v#0|pound#v#1|pound_on#v#0**
Falling back to raw glosses. These won't be displayed unless *gloss show* **raw** is checked.
Reduced to 134 entries after fallback check of copper glosses.
Found 7 gloss forms (number includes raw gloss variants) for 23 languages.
Mouse over WordNet item for sense gloss, double-click to look up base word. Double-click form for source lect sketch.

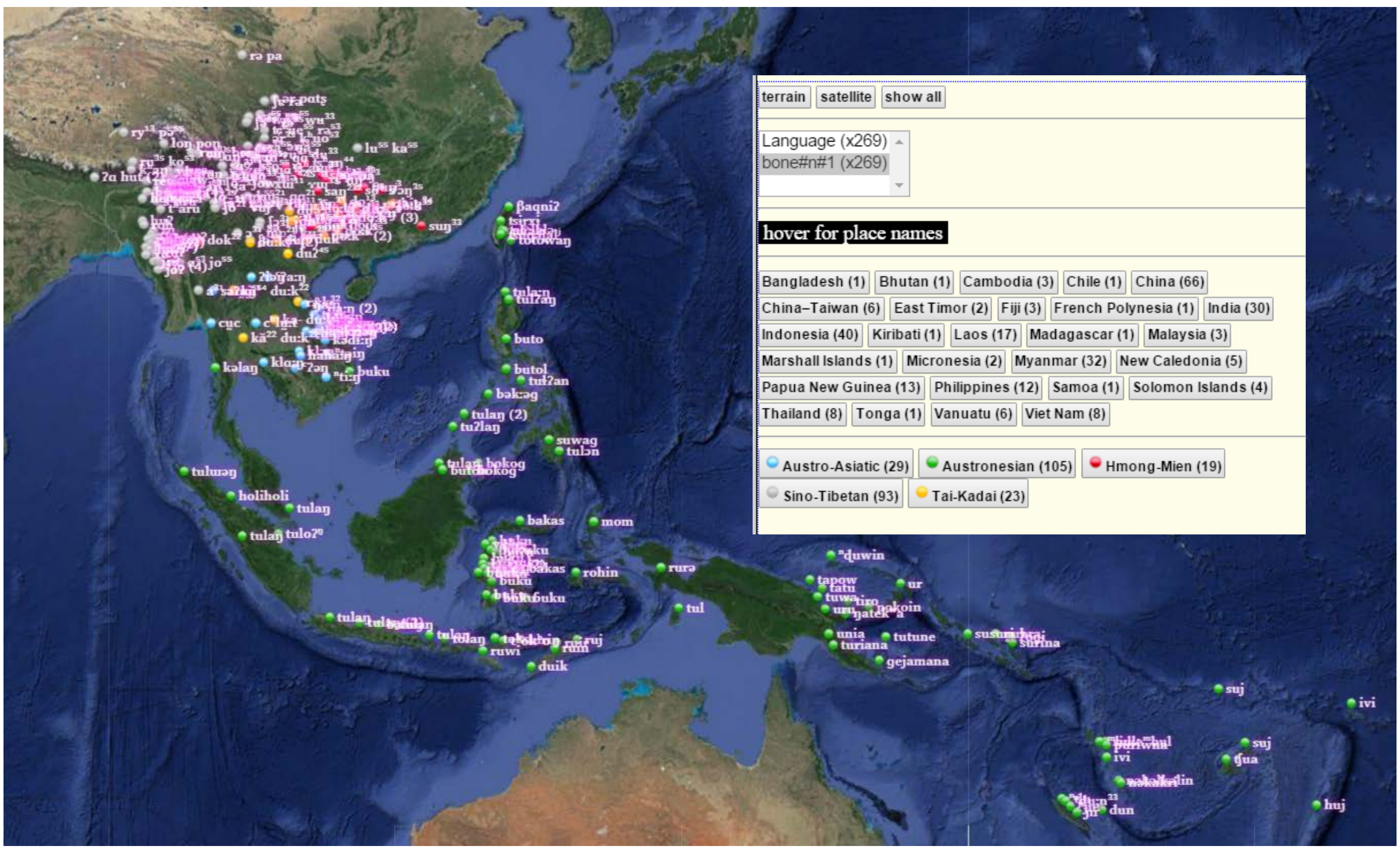
ISO	**ISO 639-3 name**	**family**	to beat, pound *hudak2008comparative* beat#v#3 | pound#v#1 (x1)	to hit the mark *zhang1999zhuang* pip#v#2 (x51)	to hit, play, etc. (in phrases) *hudak2008comparative* strike#v#1 (x9)	to pound *hudak2008comparative* pound#v#8 (x1)	to pound;to pestle *zhang1999zhuang* pestle#v#1:rice (x36)	to strike repeatedly, with a short quick motion *hudak2008comparative* strike#v#1 (x1)	to strike, as a snake *hudak2008comparative* strike#v#0 (x1)
[blt]	Tai Dam	Tai-Kadai			tok^{45}				
[khb]	Lü	Tai-Kadai			tok^{45} tok^{55}				
[nod]	Northern Thai	Tai-Kadai	tup^{454}			tam^{14}			
[nut]	Nung (Viet Nam)	Tai-Kadai			tɤk^{55}				
[shn]	Shan	Tai-Kadai			tɯk^{55}			ʃak^{55}	tɔt$^{11=21}$
[skb]	Saek	Tai-Kadai			tɯk^{454}				
[tha]	Thai	Tai-Kadai			tok^{22}				
[tts]	Northeastern Thai	Tai-Kadai			tok^{24}				
[twh]	Tai Dón	Tai-Kadai			tok^{45}				
[zch]	Central Hongshuihe Zhuang	Tai-Kadai		te:ŋ^{42} to:j^{33}			tam^{42}		
[zeh]	Eastern Hongshuihe Zhuang	Tai-Kadai		te:ŋ^{35} te:ŋ^{45} (2) to:j^{55} ʦoŋ^{53} ʦuŋ^{54}			tam^{35} tam^{45} (2)		

Frame the data
As noted above, the lexicon per se provides very little information about the language. Additional lect-specific information is usually required by downstream applications. The few choices allowed here are mainly for testing.

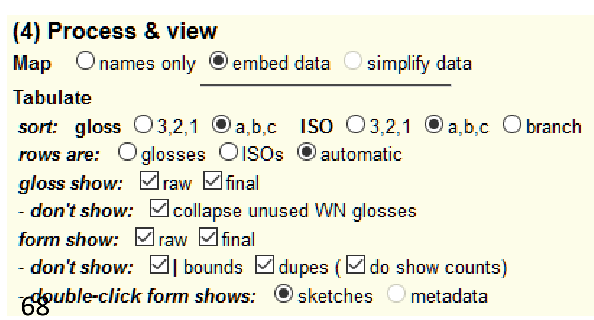
Similarly, most of the **Search** panel's controls are there to allow testing.

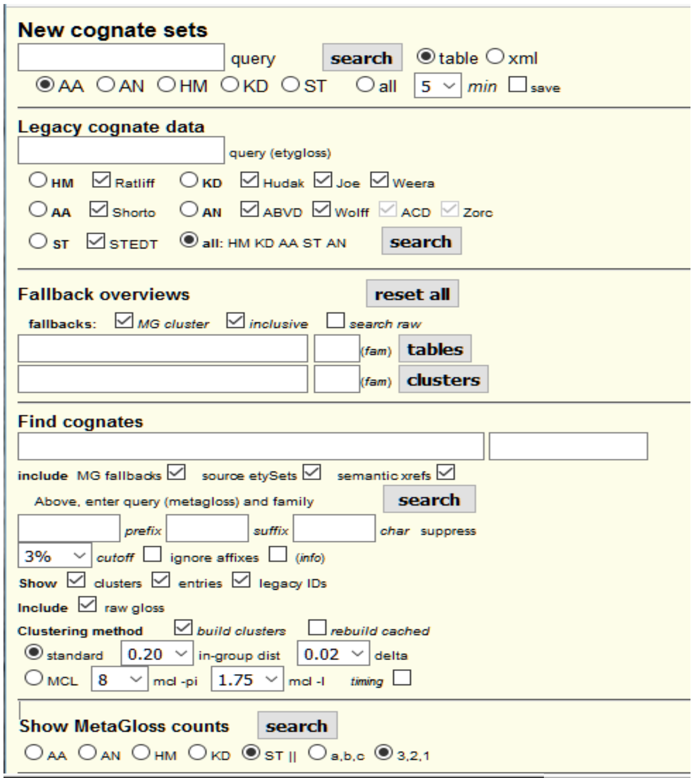


We have already seen a **table** return. A **map** view is shown below. The map control is shown as an inset. Here, we see words for **bone#n#1** drawn from all five language families. The buttons in the control allow more detailed displays by country, language family, or additional query terms. These are based on framing data that was passed through with the lexical data.



Process and view The last set of options provides more detailed control of the display. The scale of returned results varies enormously – both the lect and semantic axes may have from one to hundreds of items each – so our main concern is making very large data views manageable





A huang1992tbl (x2545)
B lsm2015chin (x468)
C lsm2015naga (x484)
D marrison1967classification (x908)

Click words below to pre-load the search interface above with partial or fully qualified WordNet / MetaGloss search terms.. Blue items are in the ABVD 210 list, asterisk items are in the LSM (similar to MSEA) list. All single-family / single word distances have been pre-calculated, as well as many (but not all) of the full MetaGloss fallback sets.*

w+n1|w+n| w || rice#n#1	731	A	175	B	365	C	130	D	61
w+p1|w+p| w || you#p#1	633	A	153	B	239	C	149	D	92
w+p1|w+p| w || we#p#1*	517	A	223	B	152	C	77	D	65
w+n3|w+n| w || paddy#n#3*	510	A	90	B	242	C	150	D	28
w+q1|w+q| w || when#q#1	473	A	47	B	239	C	149	D	38
w+v1|w+v| w || cut#v#1	471	A	172	B	147	C	75	D	77
w+a1|w+a| w || fat#a#1*	444	A	98	B	271	C	75	D	
w+a1|w+a| w || hot#a#1	443	A	49	B	252	C	103	D	39
w+n1|w+n| w || leech#n#1	442	A	35	B	241	C	144	D	22
w+n1|w+n| w || sarong#n#1	429	A	37	B	250	C	123	D	19
w+n1|w+n| w || year#n#1*	428	A	175	B	142	C	77	D	34
w+a1|w+a| w || near#a#1*	407	A	134	B	151	C	76	D	46
w+n2|w+n| w || chicken#n#2*	398	A	148	B	119	C	77	D	54
	397	A	94	B	266	C	7	D	30
	397	A	136	B	146	C	76	D	39
	391	A	99	B	144	C	76	D	72

/cogs
This page is our working tool for building cognate sets.

New cognate sets Shows sets in progress. These can be restricted by family, or by number of families represented by a given *etygloss* (the cognate set's working name).

Legacy cognate data This provides access to our database of existing comparative and proto-language reconstruction data. These help identify and provide support for new cognate sets.

Fallback overviews Raw glossing is often imprecise; even when unambiguous, semantics tend to drift over time. Thus, almost every new cognate set includes items draw from subsets with distinct raw and normalized glosses. These overview tools help us get a sense of how broadly to cast our initial net in searching for relevant cognates.

Find cognates A search for potential cognates is initiated by one or more semantic queries, usually requesting automatic inclusion of related fallback items. A phonological *distance measure* is then calculated for all returned forms, and they are clustered into potential cognate groups. The mechanisms by which distance is measured, and items are then clustered, are both highly configurable. Optimal settings are difficult to predict, and are heavily influenced by language typology.

Show MetaGloss counts Construction of cognate sets proceeds methodically through the lexicon. At this early stage, we give preference to semantics that are found in as many lects as possible. Some of the very high figures seen here an artifact of our MetaGlossing methodology – we favor *base:modifier* metaglosses, because the base generally establishes the proper cognate set.

We note in passing that the process of calculating phonological distance between *all* word pairs, and of clustering subgroups within the resultant distance tables, are both computationally quite expensive. Thus, we pre-calculate and cache huge number of distances (including all predictable fallbacks), and candidate clusters (based on a half-dozen different clustering settings).

New cognate sets This provides views of the current state of cognate set assembly. The **xml** view is saved and distributed. Below, note that the *EtyGloss* names the set, the *Refs* indicate what the original full glosses were (these are used under **find cognates**, discussed below). The individual cluster names refer to citations from the literature when possible (e.g. AA:**ash#n#1:S2034, KD:ash#n#1:W119**), and are otherwise simply numbered (**AN:ash#n#1.3**).

Overlap (EtyGloss in blue) for all families: **2** etyglosses for **6** families **33** etyglosses for **5** families **59** etyglosses for **4** families **283** etyglosses for **3** families **560** etyglosses for **2** families **2513** etyglosses for **1** families
(AA:215 AN:156 HM:518 KD:256 ST:128)

EtyGloss: AA:ash#n#1
Refs: ash#n#1 (x43) | dust#n#1 (x41) | ash#n#1:field (x3)
AA:ash#n#1:S2034 :: pʰeh | bɔh | pʰu:h | tərpo:h | buh | buh | məhəw | pɔ̤h | caʔ | pʰɔʔ | bɔ̤:h | bɑh | bɑh | bɔ:h | pʰɑʔ | bɒh | pʰɔ:w | ʔəbɔh | ʔabɔh | ʔaboh | cabuh

EtyGloss: AN:ash#n#1
Refs: ash#n#1 (x118)
AN:ash#n#1:A146.1 :: avu | abu | abɛə | fu: | aβu | abo | afu | avo | umu | e^{m}ba | kaw | gahuwej | awu | aw | aw:u | wàɔ | awuk | wahu | ahu | rehu | rɔhu | gabu | qavu | kabɔj | kəbu | refu | raβu | ləbu | zepu | ʔabóh | abuh | avuɓara | avu nu kaju | taj hapu | taj ahu | kahu | ɖaβusa: | qaβuɨiʔ
AN:ash#n#1.3 :: ^{n}dap | ^{n}dɛp^{33} | ^{n}dɛ
AN:ash#n#1.2 :: ahiᵏdesan | ahukesan | ahuklesan
AN:ash#n#1.5 :: tajaw | taʤaw
AN:ash#n#1:A146.11 :: afuafu | efuefu
AN:ash#n#1:A146.30 :: vulimolas
AN:ash#n#1:A146.8 :: feraŋa
AN:ash#n#1:A146.35 :: makola
AN:ash#n#1:A146.7 :: pe:s
AN:ash#n#1:A146.17 :: dapog
AN:ash#n#1:A146.37 :: jaj taen
AN:ash#n#1:A146.3 :: rapo rabuka

EtyGloss: HM:ash#n#1
Refs: ash#n#1 (x34)
HM:ash#n#1:R538 :: | ɕʰu^{3} | sʷaj^{3} | sɔj^{3} | ɕi^{3} | tʂʰaw^{3} | sɔ$^{3\ b}$ | tsʰu^{B} | θe^{3} | ɕʷaj^{3} | sa:j^{3} | ɕʰu^{35} | sa^{43} | sʰʲa^{33} | ɕe^{31} | θe^{53} | si^{33} | sa:j^{52} | sʷa^{53} | ɕʷaj^{53} | θʷaj^{53} | saj^{545} | ɕi^{44} | sa:j^{53} | sʷaj^{35} | ɕi^{35} | sɔj^{24} | tʂʰow^{55} | tʂʰaw^{55} | su^{13} | sʰo^{13} | sɔ^{232} | tsʰu^{55} | ɕow^{53}

EtyGloss: KD:ash#n#1
Refs: ash#n#1:plant (x38) | ash#n#1 (x18)
KD:ash#n#1:W119 :: tʰaw^{41} | taw^{31} | taw^{31} | taw^{41} | tʰaw^{52} | taw^{13} | taw^{13} | taw^{454} | taw^{44} | taw^{33} | taw^{33} | tʰaw^{44} | taw^{35} | taw^{22} | daw^{11} | taw^{213} | taw^{11} | taw^{312} | taw^{21} | taw^{42} | taw^{53} | taw^{214} | dəw^{42} | saw^{33}
KD:ash#n#1:P213 :: pʲaw^{11} | pʲaw^{11} | pʰʲaw^{11} | pʲaw^{21}

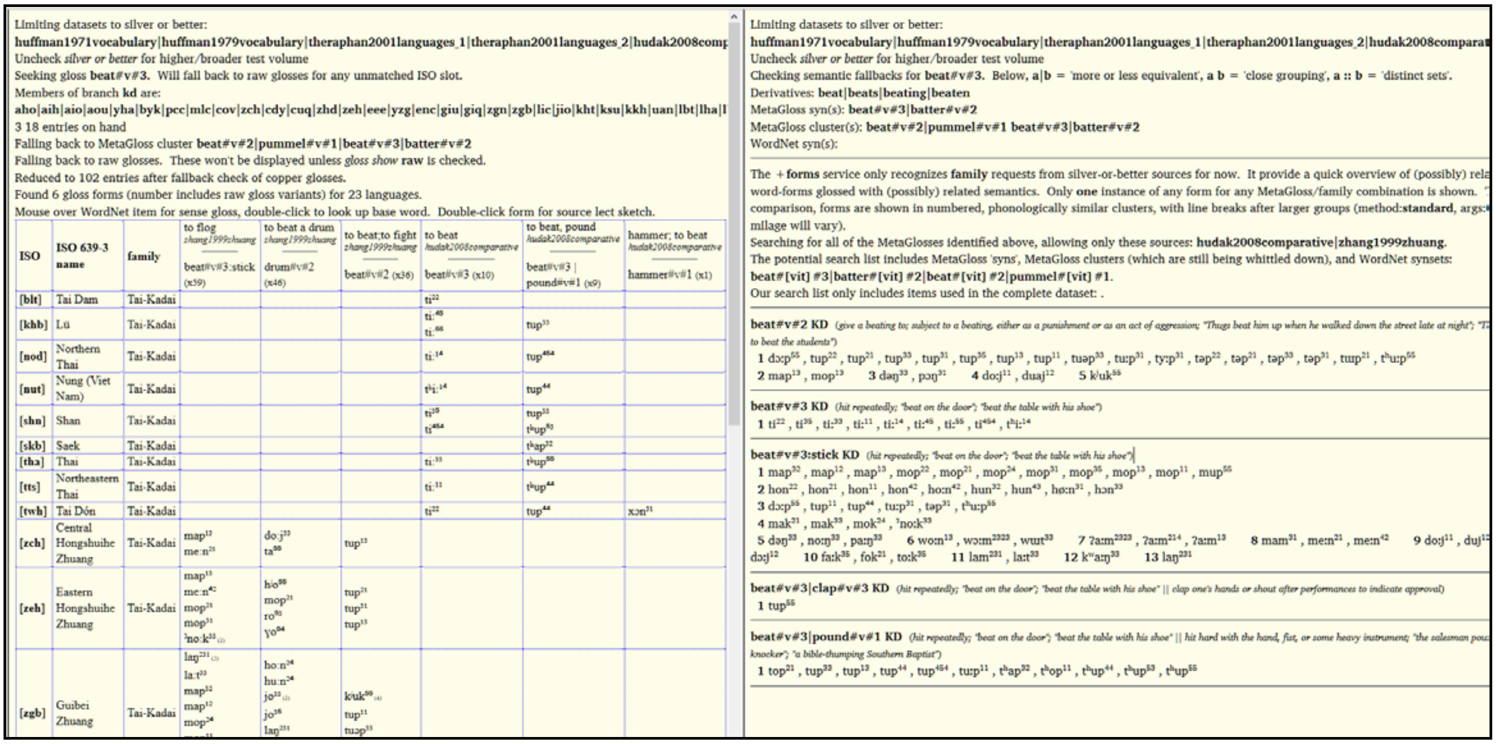
EtyGloss: ST:ash#n#1
Refs: ash#n#1 (x283) | ash#n#1:plant (x48)
ST:ash#n#1.5 :: kɨk | kuk | ^{m}kuk | ku:k | ^{m}ku:k
ST:ash#n#1.4 :: pɨŋhɨt | pɯnhɨ | pɯnhɯ | pənhət | punhət
ST:ash#n#1:M5606 :: pʰelo | tʰeklo | hɔtlʌ | tapla | tepla | təpla | tʰapla | ɑ^{31} plɑ^{53} | plɑ^{53} | luə^{35} | la^{35} | lyə^{55} | lɔ | lo | tɑp lɑ | tɑp^{51} lɑ^{4} | kʰu^{21} ɬɑ^{33} | qʰo^{11} ɬɒ^{33} | qo^{21} lɑ^{35} | lɑ tɑp
ST:ash#n#1:M374 :: go tʰal | gogtʰal | tʰalba
ST:ash#n#1:M3514 :: labu | lɑ:bu: | ^{m}but | vʊt | aot | ut | opu | ^{m}ɓut | ^{m}put | ^{m}vut | ^{m}vɨt | tɑ ^{m}vut | kʰu ^{m}vut | ku:kvɨt | ku:k ^{m}vut | vɨtpot | vutpot | vət tɑp | vɑjvət | vɑjvɨt | wɑjwut | wɑjwɨt | nivɨt | ɽɑɓɑ | lɑbu | lɑ:bu | lɑ^{4} po^{2} | lɑ^{2} bu^{2} | m̥aj pʰuʔ | majpʰu | ɓaj pʰu | bajpʰu | pʰajpʰu | majpʰø | m̥aj pʰu | wut | wit | ʔvɨt | vɨt | vut | wɨt | pu^{21} tsʰi^{35} | vuiʔ | hɑ bu | put li^{4} | bət li^{3} | bʊ^{33} tɕʰi^{33} | haj pʰu | li^{2} pʰøʔ | lipʰøʔ | ^{m}vut ʔkʰuj | ^{m}ɓut kʰuj | ^{m}put ᵏkʰuj

Legacy cognate data We rely on and refer to existing analyses whenever possible, using a separately constructed database of reconstructed proto-forms and comparative sets. In this implementation, these can be queried by semantic gloss.

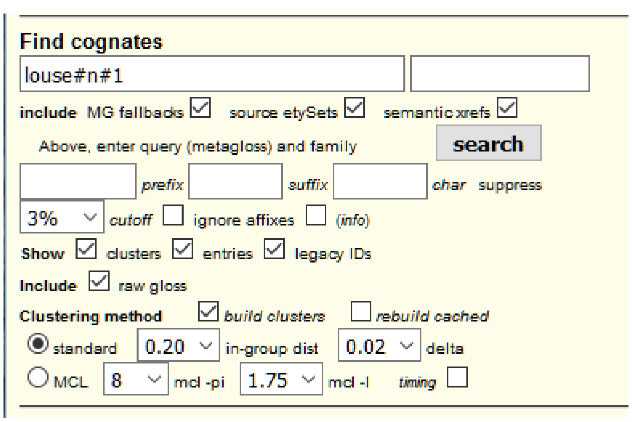
<table>
<tr><th>HM</th><th>Level</th><th>Gloss</th><th>Recon</th><th>Forms</th></tr>
<tr><td>HM:228</td><td>pHM</td><td>die#v#1</td><td>ta^{6}</td><td>tai^{6} | te^{6} | tua^{6} | tɑ^{6} | taa^{6} | ða^{G}</td></tr>
<tr><td>-</td><td>-</td><td>-</td><td>-</td><td>-</td></tr>
<tr><td>-</td><td>-</td><td>-</td><td>-</td><td>-</td></tr>
<tr><td>**KD**</td><td>**Level**</td><td>**Gloss**</td><td>**Recon**</td><td>**Forms**</td></tr>
<tr><td>KD:H346</td><td>pTai</td><td>die#v#1</td><td>--A2</td><td>haay^{1} | praay^{1} | raay^{1} | taay^{2} | taay^{1} | thaay^{1}</td></tr>
<tr><td>**KD**</td><td>**Level**</td><td>**Gloss**</td><td>**Recon**</td><td>**Forms**</td></tr>
<tr><td>KD:P704</td><td>pTai</td><td>die#v#1</td><td>*p.ta:j^{A}</td><td>ha:j^{A1} | pra:j^{A1} | pʰa:j^{A1} | ta:j^{A1} | tʰa:j^{A1}</td></tr>
<tr><td>**KD**</td><td>**Level**</td><td>**Gloss**</td><td>**Recon**</td><td>**Forms**</td></tr>
<tr><td>KD:W265</td><td>pKra</td><td>die#v#1</td><td>*pyon^{A}</td><td>pen^{A1} | phi^{A1} | phən$^{A1'}$ | puan^{c2}</td></tr>
<tr><td>-</td><td>-</td><td>-</td><td>-</td><td>-</td></tr>
<tr><td>-</td><td>-</td><td>-</td><td>-</td><td>-</td></tr>
<tr><td>**AA**</td><td>**Gloss**</td><td>**Recon**</td><td>**Forms**</td><td></td></tr>
<tr><td>AA:1266a.A</td><td>to die, be extinguished</td><td>*jap</td><td>pɔin-ɲɔp | pomɲəp-[hətə]</td><td></td></tr>
<tr><td>AA:1266a.B</td><td>to die, be extinguished</td><td>*jaap</td><td>ɲa:p | pəməɲa:p | paɲə́p | pamaɲə́p</td><td></td></tr>
<tr><td>AA:987.A</td><td>to die</td><td>*kc[ə]t</td><td>chat | kəcet | kəceet | kəciit | siit | kɛɛt | keet | chət | khchet | hacɑt | pəcɛɛt | pəceet | kəɲciit | kasiit | chɒt | sət | cə:t | cɯt | kaciat | kəcit | sɯt | cuət | cêet | həcɒt | (k)cɛ:t | (gə)sət | *cɛ:t | *kci:t</td><td></td></tr>
<tr><td>AA:1218.A</td><td>to die</td><td>*haan</td><td>phaan | há:n | phá:n | pha:n</td><td></td></tr>
<tr><td>-</td><td>-</td><td>-</td><td>-</td><td>-</td></tr>
<tr><td>-</td><td>-</td><td>-</td><td>-</td><td>-</td></tr>
<tr><td>**ST**</td><td>Recon</td><td>**Level**</td><td>**Proto-gloss group (status)**</td><td>**Morphemes** (duplicate items are suppressed). These are raw, unnormalized forms.</td></tr>
<tr><td>ST:27</td><td>*səy</td><td>PTB</td><td>DIE (ok)</td><td>'chi | 'tʃhi | chi | chu | chì | chü | ci | cə̄ | di | hai | hi | hʌi | kʰʃis | li | ntɕʰə | ri | se | sei^{35} | sej^{A} | set | se^{21} | se^{3} | se^{35} | she | shei | shi | shi^{4} | shü | shī | shi | shɛ | si | si: | sla^{31} | sle | sih | siil | sik^{31} | sil | sit | sly | sly^{3} | sly^{1} | sɪ^{3} | sɪ^{33} | sɪ^{31} | si^{11} | si: | sjər | si^{4} | si^{5} | si^{53} | si^{55} | sjhí | sjhý | sjidx | so | suh | suh: | sy | sya | syi | syid | syiy(?) | syi^{2} | syy | sẑ^{33} | sʑ^{55} | sì | sìh | sɪ́ | sü | sɪ | sɪ̆ | sɐ | sə | səj | sət | səy | səy^{1} | sə^{2} | sə^{33} | sə^{35} | sə^{13} | sə^{53} | sə^{55} | sɛ | sɤ^{13} | sɥ^{33} | sì | sī | sí | si^{55} | sɨ̈ | sɯ | sɯ˩ | sɯ̄ | sɯ^{53} | sɿ33 | sɿ^{21} | sɿ^{33} | sɿ^{11} | sɿ^{55} | sɿ^{55} | sʅ^{55} | sʰei | ši | tchhi | the^{2} | thi | thi, | thì | thí | thəy | ti | tik | tsi | tsɯ | təi | təy | tɤi | tʃhə^{1} | tʰi | tʰəy | tθe^{2} | tθe^{22} | tθi^{55} | xə | xɛʔ^{53} | xɯ^{55} | zi | śi | śi^{53} | ši | ši: | šī | ší | šī | ši | šʔ | ɕi | ɕi55 | ɕiu^{55} | ɕi^{21} | ɕi^{33} | ɕi^{31} | ɕi^{35} | ɕi^{13} | ɕi: | ɕi^{42} | ɕi^{53} | ɕi^{55} | ɕo | ɕə^{53} | ɕɛ^{53} | ɕɿ^{21} | ɦci | ʂe | ʂi^{2} | ʂi^{22} | ʂi^{21} | ʂi^{33} | ʂi^{35} | ʂo^{55} | ʂu^{53} | ʂəl | ʂə^{55} | ʂɛ | ʂɯ^{33} | ʂɯ˧ | ʂʅ^{33} | ʂʅ^{31} | ʂʅ^{55} | ʂʅ^{21} | ʂʅ^{33} | ʂʅ^{55} | ʃe^{2} | ʃer^{4} | ʃe:ɪ^{33} | ʃe^{55} | ʃi | ʃi33 | ʃih^{1} | ʃik^{31} | ʃi^{2} | ʃi^{33} | ʃi^{31} | ʃi^{33} | ʃi^{42} | ʃi^{44} | ʃi^{51} | ʃi | ʃi(?) | ʃəi^{33} | ʃɪ^{4} | ʃɿ^{33} | ʃɿ^{31} | ʃɿ^{35} | ʃɿ^{44} | ʃɿ^{44} | ʃɿ^{55} | 'c'i | θei^{53} | θe^{3} | θe^{4} | θi | θi^{55} | θí | θí | θi^{53} | θi^{55}</td></tr>
<tr><td>ST:29</td><td>*swan</td><td>PTB</td><td>DIE / SPIRIT OF DEAD / NON-VIABLE (ok)</td><td>swan | són</td></tr>
</table>

Fallback overviews
Before attempting to identify cognate set members, we usually need to get a sense of how the members are glossed, and how much semantic variation must be dealt with. Below, a search for **beat#v#3** falls back to similar semantics (**pound#v#1**), and looks for beat in raw glosses. Inclusive search allows overlap; an exclusive search only falls back when the target isn't found.

On the right, pre-clustered forms make it easy to spot probable semantic shift. Items with identical glosses are subgrouped by the similarity of their forms; this does not guarantee that they are cognates, but given that they have the exact same meaning it is highly likely. Once clusters are formed, it is fairly easy to eyeball similar groups with slightly different semantics. Again, there is no guarantee that they are cognate (or that we have found all possible cognates). However, this process helps us recognize the best starting point for the **find cognates** step.



Find cognates Below, a close-up of the menu that sets up the query. Now, in some cases data will be pre-assigned to likely cognate sets. Thus, in addition to the ordinary semantic fallback options, we are also able to expand match items to other elements of the same cognate set (even if they have different semantics – *source EtySets*), to other elements with the same divergent semantics (via *semantic xrefs*).

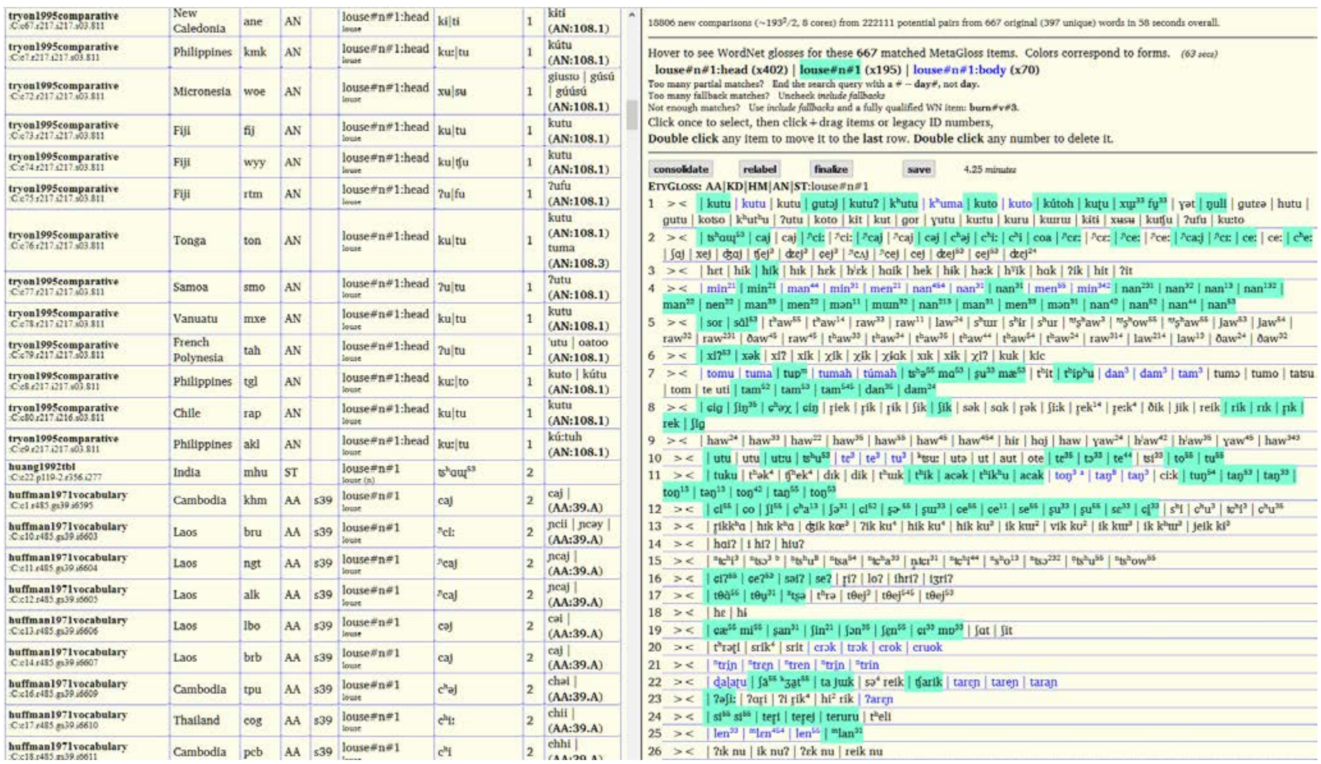


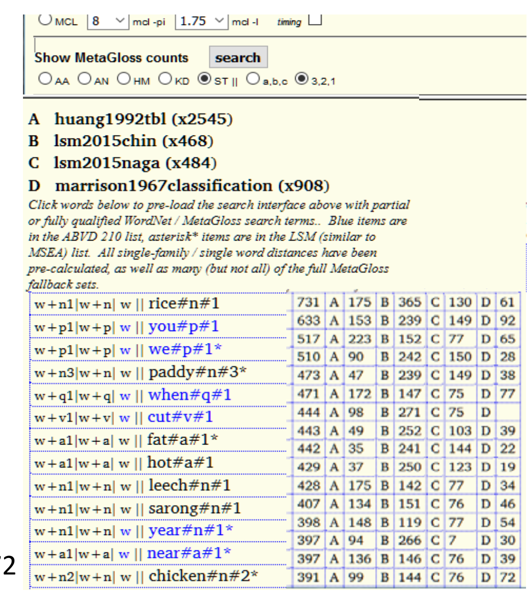
Once elements are identified, we assess their phonological distance, and cluster the closest elements. Now, depending on language typology, the cognate *morpheme* might not typically be a free *lexeme* – for example, some languages might tie it to a class term that means "fruit" or "animal". In order to create more accurate distance measures, we provide mechanisms for suppressing part of the returned forms, either by specifying an affix to ignore, or by assessing each lect's complete word list, and inferring likely affixes.

Clustering methods are also configurable. This implementation allows two types: a bottom-up *agglomerative tree-building* approach that is bounded by the maximum distance between any two items, and *Markov Chain clustering*, which can be more effective for properly clustering items from relatively continuous dialect chains.

Below, we see the result of a search in all five families for **louse#n#1** and its MetaGloss fallbacks. On the left each item is shown with source and language information, the raw gloss, the proposed cluster, and any additional information that could be derived from the **legacy cognate data** discussed above.

On the right, each alternative semantic is colored differently; this is helpful assessing likely cognate status. It will not be obvious, but in this case each of the clusters on the right naturally falls into a language family-specific grouping: 1/AN, 2/AN, 3/HM, etc.





Show MetaGloss counts This control lets us look at the distribution of semantic items within the database. Not every source has the same coverage; this view foregrounds items with broad representation. It was also helpful in refining MetaGloss assignments – unexpected gaps in semantics that were due to inconsistent choice of specific MetaGlosses.

This is primarily a production tool, intended to let us survey data as quickly as possible. Thus, all of the non-numeric values are actionable, usually to pre-load other parts of the menu.